Speech, Audio, Image and Video Technologies (SAIVT)
School of Electrical Engineering and Computer Science
Computational Intelligence and Signal Processing Discipline

Sequential Decision Fusion of Multibiometrics applied to Text-Dependent Speaker Verification for Controlled Errors

Vishnu Priya Nallagatla

Bachelor of Engineering (Computer Science)

SUBMITTED AS A REQUIREMENT OF

THE DEGREE OF DOCTOR OF PHILOSOPHY

AT

QUEENSLAND UNIVERSITY OF TECHNOLOGY

BRISBANE, QUEENSLAND

MAY 2012

Keywords

Multi-Instance Fusion, Multi-Sample Fusion, Verification Error Trade-off, Sequential Decision Fusion, Correlation Modelling, Bahadur-Lazarsfeld Expansion, Favourable Statistical Dependence, Classifier Selection, Sequential Error Ratio

Abstract

Reliability of the performance of biometric identity verification systems remains a significant challenge. Individual biometric samples of the same person (identity class) are not identical at each presentation and performance degradation arises from *intra-class variability* and *inter-class similarity*. These limitations lead to false accepts and false rejects that are dependent. It is therefore difficult to reduce the rate of one type of error without increasing the other. The focus of this dissertation is to investigate a method based on classifier fusion techniques to better control the trade-off between the verification errors using text-dependent speaker verification as the test platform.

A sequential classifier fusion architecture that integrates multi-instance and multi-sample fusion schemes is proposed. This fusion method enables a controlled trade-off between false alarms and false rejects. For statistically independent classifier decisions, analytical expressions for each type of verification error are derived using base classifier performances. As this assumption may not be always valid, these expressions are modified to incorporate the correlation between *statistically dependent* decisions from clients and impostors. The architecture is empirically evaluated by applying the proposed architecture for *text dependent speaker verification* using the Hidden Markov Model based digit dependent speaker models in each stage with multiple attempts for each digit utterance. The trade-off between the verification errors is controlled using the parameters, number of decision stages (instances) and the number of attempts at each decision stage (samples), fine-tuned on evaluation/tune set. The statistical validation of the derived expressions for error estimates is evaluated on test data.

The performance of the sequential method is further demonstrated to depend on the order of the combination of digits (instances) and the nature of repetitive attempts (samples). The false rejection and false acceptance rates for proposed fusion are estimated using the base classifier performances, the variance in correlation between classifier decisions and the sequence of classifiers with favourable dependence selected using the *'Sequential Error Ratio'* criteria. The error rates are better estimated by incorporating *user-dependent* (such as speaker-dependent thresholds and speaker-specific digit combinations) and *class-dependent* (such as client-impostor dependent favourable combinations and class-error based threshold estimation) information.

The proposed architecture is desirable in most of the speaker verification applications such as remote authentication, telephone and internet shopping applications. The tuning of parameters - the number of instances and samples - serve both the security and user convenience requirements of speaker-specific verification. The architecture investigated here is applicable to verification using other biometric modalities such as handwriting, fingerprints and key strokes.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

AVICAR	Audio-Visual speech In a CAR
BLE	Bahadur-Lazarsfeld Expansion
CSLU	Centre for Spoken Language Understanding
DET	Detection Error Trade-Off
EER	Equal Error Rate
FAR	False Acceptance Rate
FRR	False Rejection Rate
GMM	Gaussian Mixture Models
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
LPCC	Linear Predictive Cepstral Coefficient
MAP	Maximum A Posteriori
MCE	Minimum Combination Error
MFCC	Mel Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
PSOLA	Pitch Synchronous Overlap and Add
ROC	Receiver Operating Characteristic
SER	Sequential Error Ratio
SPRT	Sequential Probability Ratio Test
TDSV	Text Dependent Speaker Verification
TER	Total Error Rate
UBM	Universal Background Model
VTLN	Vocal Tract Length Normalization

Authorship

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signed: QUT Verified Signature

Date: 25/01/13

Acknowledgements

First and foremost, I would like to thank my family for their moral support and constant encouragement which was crucial for the completion of this dissertation. Special thanks to my mother for including me in her every prayer and to my father for always being there for me. Thanks to my elder sister and brother-in-law for accepting nothing less than completion from me.

I would like to offer my sincerest gratitude to my supervisor, Professor Vinod Chandran, who has supported me throughout my dissertation with his patience and knowledge while allowing me the room to work in my own way. I am also grateful to Professor Sridha Sridharan for his suggestions after the seminars and providing me with access to the SAIVT lab resources. I would also like to thank Anthony Nguyen and everyone within the Speech, Audio, Image and Video Technologies (SAVIT) laboratory. Thanks to Professor Prasad Yarlagadda for helping me with the opportunity to study at QUT.

I would like to acknowledge the contributions of Dr. Eddie Wong and D. Lawson in the QUT Vocal Access Speaker Verification software package used for testing in this dissertation. The acknowledgement extends to Dr. David Suendermann for providing access to Voice Conversion Matlab® Toolkit.

My friends - thank you for not being mad at me of long gaps in communication during the write up of the dissertation. Thanks to my family and friends here in Australia for their support. Lastly, thanks to Manoj for his patience with my mood swings and love to my nephews (Jyothin, Lokajith and Lalat) for helping to keep me sane.

Chapter 1
Introduction

This PhD Dissertation is focused on the identity verification using multiple sources of information from a single biometric characteristic. In particular, this dissertation explores the benefits of sequential fusion of biometric information in identity verification for better control over the trade-off between verification error rates - false accepts and false rejects. The architecture is empirically evaluated by applying the proposed architecture for *text dependent speaker verification* using Hidden Markov Model based digit dependent speaker models.

The increasing importance of reliable personal identity recognition by automatic means has resulted in establishment of the technological area known as biometrics [1]. The term *'biometrics'* refers to the automatic recognition of an individual based on physiological and/or behavioral characteristics (e.g., fingerprints, face, iris, voice, signature, etc.). Biometric technologies have been used in not only government, legal or forensic operations but also in a large number of civilian applications where the fundamental task is to establish the identity of individuals. Some of the large-scale biometric systems include the Integrated Automated Fingerprint Identification System (IAFIS) of the FBI, the US-VISIT IDENT program, the Schiphol Privium scheme at Amsterdam's Schiphol Airport and the finger scanning system at Disney World, Orlando [2].

This chapter presents the basic architecture of biometric verification systems, including common performance measures and the reasons for combining multiple biometric sources. The detailed information on the introductory topics in biometrics is provided by Jain et al. [1, 3]. The three main design issues of multibiometric system - classifier fusion architecture, classifier correlation and classifier selection - are also outlined. The chapter also explains the motivation and outline of the dissertation. The last sections state the research objectives and original contributions from this work.

1.1 Biometric System

Biometric recognition, or simply biometrics, is a natural and reliable solution to the problem of determining the identity of an individual. A biometric recognition system

essentially uses either behavioural or physiological biometrics, modeled by means of pattern recognition and statistical methods. Physiological biometrics are based on a person's physical characteristics which are assumed to be relatively unchanging (passive) over time, such as fingerprints, iris patterns, retina patterns, facial features, palm prints, or hand geometry. On the other hand, behavioral characteristics, such as signature or voice, are dependent to some degree on the individual's state of mind and acquired over time (active). Certain biometric characteristics, e.g., *voice biometrics* [4], are considered the combination of physiological and behavioral characteristics. Voice from an individual depends on both the physical features such as vibrations of vocal cords and vocal tract shape and on behavioral features such as the state of mind of the person who speaks.

Biometric systems offer several advantages over traditional security methods that are based on something that you know (knowledge information such as a secret password or PIN, which can be shared, forgotten or copied) or something that you have (a physical object). As biometrics cannot be forgotten or lost and are difficult to forge without sophisticated methods, the solutions provided by biometrics offer high security. The performance of the biometric system is enhanced by the combination of possession and/or knowledge information with biometrics. An important issue in designing a practical biometric system is to determine the method of recognition for an individual. Depending on the application context, a biometric recognition system can perform either verification or identification of an individual [3]:

- A verification system recognizes a person's identity by comparing the acquired biometric characteristic with his/her previously enrolled biometric reference model pre-stored in the system. It conducts one-to-one comparison to confirm whether the identity claim of the individual is true. A verification system either rejects or accepts the submitted identity claim.
- An identification system recognizes an individual by searching for a similarity in the entire enrolment model database. It conducts one-to-many comparisons to establish if the individual is present in the database and if so, returns the identifier of the enrolment reference that matched. In this context, the system establishes an identity (or determines if the individual is not enrolled in the system database) without the individual having to claim the identity.

This dissertation focuses on biometric verification. Figure 1.1 shows the two modes of operation in a biometric verification system, i.e., enrolment/training and verification. The biometric system is essentially a pattern recognition system that makes use of the data

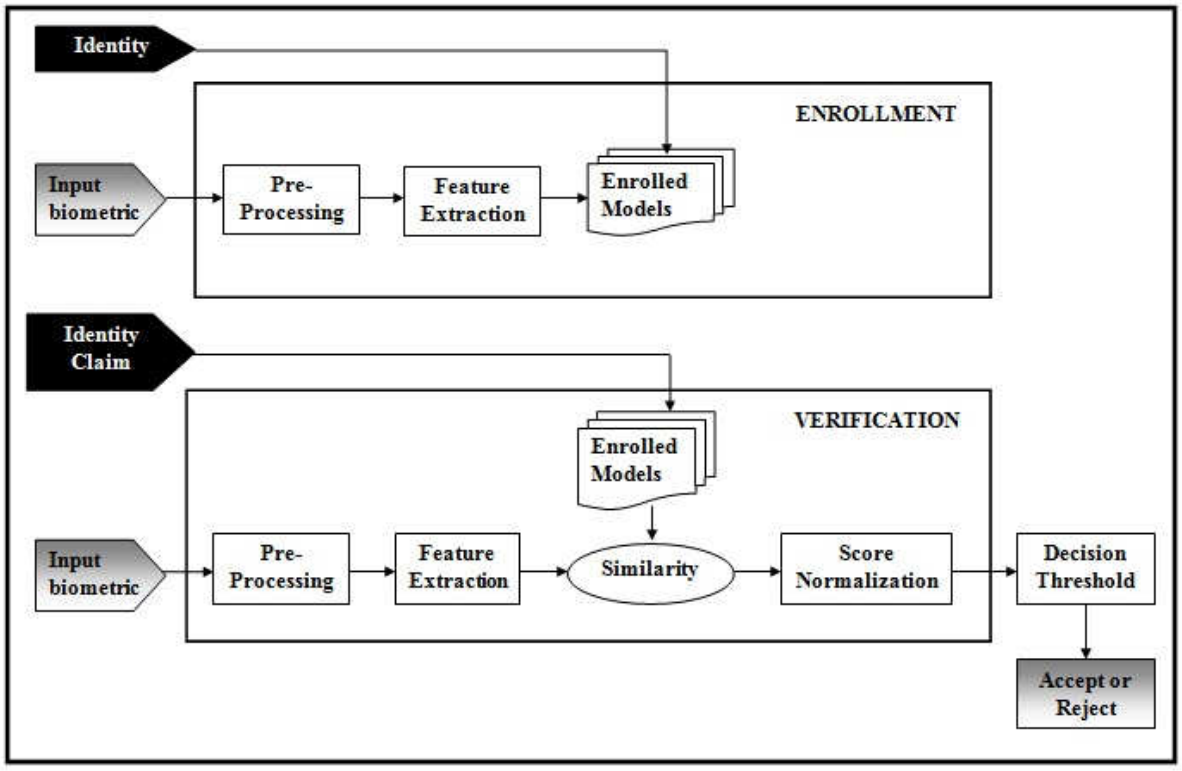

Figure 1.1 Modes of operation in a verification system (a) Enrolment and (b) Verification

acquisition and prepossessing module, feature extraction module, matching module and decision-making module. The system acquires biometric data from an individual and processes this information to extract a set of salient features. The extracted feature sets are used to create models during enrolment. The feature sets extracted during verification are then compared against the claimant model (stored in the database) and a similarity/matching score is produced. If the score is higher than a decision threshold, then the claim is accepted, otherwise rejected.

The objective in biometric verification is to classify the input biometric signals into two classes, either client or impostor. It is possible that impostors will exploit the known information regarding the biometric characteristics of a client and thus verification performance degrades. As a result, two kinds of impostors are usually considered, namely: i) casual impostors (producing random utterances in case of speaker recognition) when no information about the target user is known and ii) real impostors (producing mimicked or voice transformed utterances in the case of speaker recognition) when some information regarding the biometric characteristic being forged is used. Most of the standard databases for biometrics have information of several individuals and so the verification tests for an individual considers information from others as casual impostor data.

Irrespective of the nature of modality, either physiological or behavioral, one biometric may not fulfil the application requirements in terms of the properties such as universality, distinctiveness, performance, circumvention and system cost [5]. The focus of this research is the sub-area of ***voice biometrics*** which is considered a fairly natural technique for everyday transactions [6]. The cost associated with a voice biometric system is *lower* than other biometrics as no special hardware, but standard phones or microphones, is required for recording speaker utterances. In addition, voice biometrics is one of the most preferred biometrics for remote authentication that does not need speaker's physical presence.

If recognition is based on speech that is selected by the speaker and not known ahead of time, the recognition system is *text-independent*. When the system is trained on *a* particular utterance and later a decision is made on the same utterance then the recognition system is *text-dependent*. The text-dependent speaker recognition system with random or time-variant text will be hard to spoof thereby satisfying the requirement of *circumvention* [7]. The *performance* of voice biometrics, however, depends on a number of factors such as intra-speaker variability, background noise, quality of the channel used for recording and communication. As with voice, each of the other biometrics has its own strengths and weaknesses for the application properties [5]. The applicability of a specific biometric characteristic depends heavily on the requirements of the application and no single characteristic is shown to be *'optimal'* in the sense that it out-performs all the other biometrics in different operational environments.

1.2 Performance of a Biometric System

The major concern in biometric verification is its accuracy. One general problem of biometrics is that the individual biometric samples of the same person are not identical for each presentation. This *intra-class variability* is caused by several reasons such as different environments, changing sensors or even natural biometric variability. *Inter-class similarity* occurs owing to common characteristics of the same biometric modality between different persons. These limitations may lead to misclassification of the verification claims resulting in false accepts and false rejects. False rejection refers to the likelihood of an authorised user being wrongly rejected by the system. False acceptance refers to the likelihood of an impostor being wrongly accepted by the system. The two errors are complementary - when one type of error is reduced by varying the threshold, the other error rate automatically

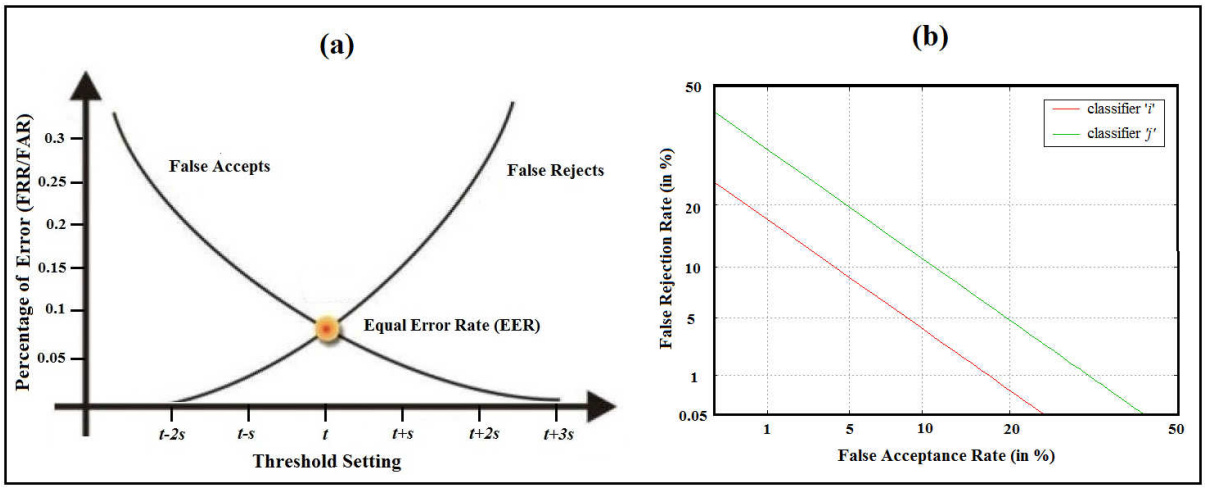

Figure 1.2 The performance of a biometric system summarized using (a) FRR and FAR curves against decision threshold and (b) DET curve that plots FRR against FAR in the normal deviate scale

increases. The balance between these error rates is found with a decision threshold that can be specified either to reduce the risk of FAR, or to reduce the risk of FRR.

In an ideal-world scenario, the curves showing instances of false accepts and false rejects meet at a point of zero errors at some threshold setting of '*t*'. In this case, the adjusting of threshold setting in either direction (lowering or increasing threshold) would be of little point because a change in the threshold '*t*' here would either increase false accepts or false rejects. However, in the real world, things are not so clear and typically, the biometric system will not meet this ideal curve with the two curves intersecting at a point above zero errors (e.g., fig 1.2(a)). In these situations, it is difficult to achieve zero false accepts and zero false rejects, but the performance biased towards increased false rejects or false accepts is obtained by adjusting the threshold setting. For example, some high-security applications, such as ATM, the access control of nuclear power stations and exchequers [4, 5], require extremely low false acceptance rate from the biometric system. In these applications, the rejection of an authentic person might just be troublesome. While, if an impostor is mistakenly accepted, it may be a disaster. Under these conditions, the threshold could be tightened (e.g., threshold *'t+2s'* in fig. 1.2 (a)) to lower the number of false accepts. However, because of the trade-off between the biometric error rates, the increase in the threshold might result in an increase in FRR (in fig. 1.2(a) FRR increases by 0.25%).

When the threshold is loosened the number of false rejects are reduced (e.g., the setting of threshold *'t-s'* results in low FRR with an increase in false accepts). For example,

commercial service providers (such as retail stores or bank) are likely to take a much more liberal view of permitting impostors to commit fraud in their systems rather than losing transactions from an authentic user, causing user inconvenience. Hence, the trade-off between FAR and FRR translates into a trade-off between security and convenience.

The error rates at various values of threshold '*t*' can be summarized using a Detection Error Trade-off (DET) curve [8] that plots FRR against the FAR at various thresholds on a normal deviate scale and interpolates between these points (fig.1.2(b)). In addition, the single-valued measure known as Equal Error Rate (EER) summarizes the performance of a biometric system. The EER refers to that point in a DET curve where the FAR equals the FRR and a lower EER value indicates better performance. When designing a biometric verification system, the first question to decide on is: *"Is security of prime concern, or is convenience the real issue for the application?"* Another consideration in the design of a biometric system is the tradeoff between cost and security of the verification. The cost of the system can be measured using the verification errors, higher the FRR or/and FAR, the more

Table 1.1 State-of-the-art error rates associated with fingerprint, face and voice biometric systems [9, 10].

Biometric Trait	**Test**	**Test Conditions**	**FRR**	**FAR**
Fingerprint	FpVTE 2003	US Government operational data	0.6%	0.01%
Face	FRVT 2006	Controlled illumination, high resolution	0.8-1.6%	0.1%
Iris	ICE 2006	Controlled Illumination, broad quality range	1.1-1.4%	0.1%
Voice	NIST 2000	Text Dependent	10-20%	2-5%
	NIST 2004	Text Independent, multi-lingual, operational data	5-10%	2-5%
	NIST 2008	Text Independent	12%	0.1%
	NIST 2010	Text Independent, eight two-channel conversational telephone test speech	9%	0.15%

expensive the application, as more subjects are incorrectly authenticated. Based on this requirement an operating point/threshold is selected to obtain the desired level of error rates.

Biometric systems deployed in a number of real-world applications have higher error rates associated with them. Table 1.1 provides a good estimate of the error rates of the state-of-the-art biometric (fingerprint, face, voice and iris) systems obtained through various technology evaluation tests. These error rates are obtained by the testing of competing algorithms on common databases, but the evaluation is dependent on a number of test conditions such as the sensor used, the acquisition protocol, the number of the subjects involved and the time lapse between successive biometric acquisitions [1]. Although commercial biometric systems have comparably lower error rates than in table 1.1, the methods still cannot identify individuals with complete accuracy.

Lower error rates for a biometric system are obtained with an appropriate design. The factors that contribute to the complexity of system design include qualitative measures such as usability and quantitative parameters like accuracy. Usability of a biometric system depends on the choice of the biometric characteristic, design of the user interface and sensor quality. The designed system should provide an interactive interface that alleviates the user's ability in providing reliable samples for verification. Accuracy requirements for a biometric system are application dependent and are affected by factors such as [1] :

- Noise in acquired data due to imperfect or variable acquisition conditions resulting in an individual being incorrectly rejected by the system.
- Non-universality of a biometric modality due to individuals with non-meaningful data requiring an exception procedure to handle with them.
- Lack of distinctiveness of the biometric modality due to an implicit upper bound in the recognition accuracy.
- Spoof attacks by means of imitation of behavioural characteristics (voice, signature, etc.) or synthetic reproductions of physiological characteristics (e.g. fingerprint or iris) resulting in impostors being incorrectly accepted.

Some of these limitations of a single biometric verification system are addressed by designing new sensors to acquire reliable biometric characteristics, adapting robust and efficient matching algorithms. However, the design of a biometric system that combines evidences from multiple biometric sources compensates the limitations of the individual sources. Such systems, known as *multibiometric systems*, are more accurate due to the integration of information from multiple sources of biometrics.

1.3 Multibiometric System

The integration approach for biometric verification is employed when it becomes increasingly difficult to achieve significantly better performance using a single biometric. Biometrics when combined with non-biometric information such as possession or knowledge-based schemes provides multi-factor authentication. Although these systems reduce the false acceptances, the problems inherent in the possession- and knowledge-based techniques are re-introduced. Therefore, the alternative is to combine multiple biometrics themselves.

Multibiometric systems combine information from multiple sources of a single or multiple biometric characteristics. Such systems are expected to be more reliable with some advantages over uni-biometric system as different biometric sources usually compensate for the inherent limitations of the other sources [11]. The use of multiple biometrics can significantly improve the overall accuracy of a biometric system and makes more resistant to spoof attacks, as it is difficult for an intruder to spoof multiple biometrics simultaneously.

Multibiometric systems also have a few disadvantages when compared to single biometric systems. They are more expensive and require additional resources for computation and storage than single biometric systems. These systems generally require additional time for user enrolment causing some user inconvenience. Io improve the accuracy of the multibiometric verification system, the multiple sources of information should be combined using an appropriate design that, in general, depends on application requirements. The major issues to be addressed in the design of the multibiometric system are [1]:

- *Sources of Multiple Biometrics* that includes information from multiple sensors, multiple representations and matching algorithms, multiple samples or multiple instances of a biometric characteristic and multiple biometric characteristics. The choice of biometric sources used for the design of multibiometric system depends on the requirements of the application.
- The architecture for *acquisition and processing* of information from multiple sources is either serial (cascade or sequential), parallel or hierarchical. Depending on the application scenario, the data from multiple sources is acquired/processed either simultaneously (parallel) or at different instances (serial).
- The *choice of information* used for fusion has a significant impact on the performance of the system. Depending on the type of information, the fusion scheme is classified into

sensor level, feature level, score level and decision level fusion schemes where the information is in the form of raw samples, feature sets, match scores or decision outputs respectively.

- Given the type of information from multiple sources, the challenge is to find the *optimal fusion techniques* that can be admissible for an application.

Each of these design issues is discussed in detail in chapter 2. Although multibiometric systems are shown to improve performance over uni-biometric systems, it is difficult to predict the optimal design factors - sources of biometric information, mode of operation or information and fusion methodology - relevant for a particular application based on performance alone [12-14]. Additional factors such as system cost, user convenience, scalability also play a significant role in selecting the design of the multibiometric verification system based on application specifications.

Multibiometric systems can be integrated at multi-classifier level and thus be considered as a conventional fusion problem where information from different biometrics is combined. Biometric fusion has been empirically shown to improve the accuracy of biometric verification and overcome the weakness of individual classifiers [11, 15] using appropriate combination methods. Kittler et al. [16] compared various classifier combination schemes experimentally and some of them were shown to consistently outperform a single best classifier. Most fusion solutions aim for the highest multibiometric accuracy by minimizing one type of error for a fixed value of another verification error. Nandakumar et al. [17] proposed the use of a fusion approach based on the likelihood ratio in the context of biometrics that directly minimizes the false reject rate (FRR) at the specified values of false accept rate (FAR). The multibiometric fusion architecture proposed in this dissertation aims to reduce both verification error rates - false accepts and false rejects - simultaneously with the trade-off controlled using the choice and number of biometrics used for verification.

1.4 Motivation of the Dissertation

The dissertation proposes a novel sequential decision fusion strategy that enables a better control over the trade-off between false rejection and false acceptance rates of a verification system. The principle of fusion strategy is to consider serial combination of uni-modal systems to reduce the false accepts, where the decision at each system is made sequentially from additional biometrics to reduce the false rejects. The tuning between the number of uni-modal systems and the limit on repeated tries in sequential fusion enables to

improve the performance of the multibiometric verification system. The architecture is applicable to any biometric modality and thus the proposed fusion scheme is applied to text-dependent speaker verification as a test platform. The research carried out in this dissertation is motivated by the following observations from literature:

- The first observation comes from the contributions by Takahashi et al. [18] and Vildjiounaite et al. [19] where serial fusion approaches were used for multi-modal biometric systems. Takahashi et al. [18] applied sequential probability ratio test (SPRT) method for multi-modal decision fusion to minimize the average number of inputs. This method is shown to enable quantitative control of accuracy with FAR limited to a specified value independent of the input order. Vildjiounaite et al. [19] proposed the cascading of unobtrusive biometrics with more reliable biometrics in such a way that later ones are required only if unobtrusive verification fails. Since performance of unobtrusive biometrics is not sufficiently high for achieving low FRR and FAR, the first biometric selected falls with reasonably low desired FAR. In the next stage, a more reliable biometric is used to complement the unobtrusive verification. The above two architectures can be integrated such that multiple classifiers are combined in serial architecture with sequential fusion of biometrics at each classifier level. Such architecture has the possibility of reducing both verification error rates simultaneously.
- The second observation is from the working principle of *"Vocal Access Text-Dependent Speaker Verification"* system developed in QUT under the supervision of Prof. Sridha Sridharan. The idea is to perform speaker verification using 16 prompted isolated digits. If individual digits at each prompt are recognised correctly, verification is performed using a single score threshold on the combined score from individual digits. If any digit is not recognised, the user is allowed another attempt for the same digit.

Utterance length has a significant effect on overall system performance [20]. Surendran [21] presented the use of sequential fusion methodology on smaller chunks of data for faster verification. Experimental results demonstrated that the decisions made using about seven digits per utterance were as reliable as using a fixed length of 10 digits resulting in about 30% savings in computational cost. The principle idea of Vocal Access TDSV when modified to include the use of individual digits (smaller length utterances) provides a reliable verification decision without having to acquire and process all digits in the prompt. If the decisions are combined such that a rejection is made at any digit level but acceptance only at final stage (if and only if accepted by all digits in sequence), the architecture reduces the number of false accepts.

The increase in false rejects because of the serial architecture is reduced by allowing multiple attempts at each digit verification stage. The combination of multiple attempts in turn increases the false accepts. However, Kashi and Nelson [22] have demonstrated that the fusion of multiple signature samples results in an increase in FAR that is small compared to the reduction in FRR. Therefore, the use of sequential fusion method for combination of decisions from multiple digits with multiple utterances at each digit level improves the overall performance of TDSV.

- The third observation is the analysis of a fusion method similar to the architecture explained in second observation. Chandran and Nguyen [23] pointed out with expressions for verification errors that the significant decrease in false acceptance rate for the combination of classifiers need not be actually traded off with false rejection rates. The trade-off is the increase in total time taken for allowing multiple attempts at each classifier. Expressions for controlled trade-off of errors were derived in this work for the case of statistically independent decisions.

Based on the above observations, a multibiometric architecture is proposed that improves the fusion performance with control over the trade-off between false rejects and false accepts for biometric verification. Also considered is the idea of multifactor verification that combines biometrics with knowledge-based verification mechanism. The architecture investigated here is directly applicable to speaker verification from spoken digit strings such as credit card numbers in telephone or voice over internet protocol based applications. Although validations presented in the dissertation are limited to use of multibiometrics (multiple instances and multiple samples) from voice biometric, the design architecture is applicable to other biometric characteristics such as handwriting and fingerprint.

1.5 Outline of the Dissertation

The main objectives of the PhD dissertation are as follows:

1) studying the performance of multibiometric verification system that incorporates the sequential fusion of decisions from multiple instances and multiple samples, with consideration to the nature of repetitive samples (adaptive and random).
2) investigating the trade-off in verification error rates, i.e., false acceptance and false rejection rates, for the proposed multi-instance and multi-sample fusion architecture under the assumption of independence between the classifier decisions

3) investigating the effects of modelling correlation between the classifier decisions on the verification error rates for the proposed sequential fusion architecture
4) determining a best classifier selection method for a subset of classifiers with optimal performance for sequential fusion
5) investigating the effect of incorporating *user-dependent* and *class-dependent* information on the number of false rejects and false accepts for the proposed sequential fusion architecture

The introduction to the topics of biometrics and multibiometrics are presented in chapter 1 along with the motivation, outline and contributions of this PhD dissertation. The major multibiometric design issues are detailed in chapter 2. This chapter presents the basic architectures for multi-instance and multi-sample fusion schemes with an explanation on sequential architecture that integrates these two fusion schemes. The chapter also provides the theoretical analysis of verification error rates under the assumption of statistical independence between classifier decisions.

As the proposed architecture considers the fusion of information at decision level, it is applicable to verification based on any of the biometric modalities such as voice, fingerprints, keystroke dynamics, handwriting samples. The theoretical analysis on verification error rates is experimentally evaluated by applying the proposed architecture for speaker verification, where the identity claim is verified based on his/her voice characteristics. The architecture considers the use of digits data as multiple biometric instances and multiple presentations of the same digit as multiple biometric samples in case of voice biometrics. The software for *'Vocal Access Text-Dependent Speaker Verification'* system is modified to obtain decisions from each digit and evaluate the expressions developed for the biometric error rates.

Chapter 3 introduces the basic architecture of speaker verification with an explanation on modelling techniques and issues related to verification based on text dependence. This chapter also provides the design and protocol used for experimental analysis of multi-instance and multi-sample fusion schemes. Chapter 4 presents the experimental results of multi-instance and multi-sample fusion schemes for text-dependent speaker verification using Hidden Markov Models (HMM) based digit dependent speaker models. The empirical results for sequential integration of multi-instance and multi-sample fusion schemes (that controls the trade-off between verification errors) and the statistical validation for comparison of these experimental error rates with theoretical error rates calculated under independence

assumption are among the main contributions of this PhD dissertation, therefore, presented in detail with consideration to *user-dependent* and *class-dependent* measures.

Although the assumption of statistical independence holds for some applications in multi-modal biometric fusion, it is often unrealistic for other multibiometrics such as multi-instance and multi-sample fusion schemes. In chapter 5, the independence assumption is relaxed to consider the dependence relationship between the classifier decisions for further refinement of the statistical analysis of proposed fusion. This chapter introduces the statistical dependence between decisions and then its effect on multi-instance and multi-sample fusion schemes. The expressions for biometric error rates are modified to incorporate correlation between the verification decisions and are empirically evaluated using a text-dependent speaker verification test platform. The chapter also presents the theoretical and analytical analysis for determining favourable/unfavourable dependence conditions for the proposed architecture.

The design of the proposed multibiometric system architecture is optimised by selecting a subset of classifiers with optimal performance. Chapter 6 introduces the classifier selection methods and most commonly used criteria in biometric literature for selection of classifiers with optimal performance. A new criterion - *'sequential error ratio'* is proposed which is specifically tuned to the characteristics of the proposed fusion scheme. The empirical comparison of existing selection criteria with the proposed measure is presented in this chapter for fusion of multiple instances with and without repetition of samples. Chapter 7 concludes the dissertation by summarizing the contributions and outlining future research directions.

1.6 Areas of contributions

The error rates for the state-of-the-art uni-biometric systems obtained through various algorithms and methods still cannot identify individuals with complete accuracy (table 1.1). Accuracy is improved by simply bringing the fundamental task of finding the best features and best classifiers to a different level i.e., the best set of classifiers and then the best combination method [24]. The literature on classifier combination and multibiometrics is vast [1, 24, 25] and much work exists on applying classifier combination/fusion to speaker verification (e.g. [26-30]). The dissertation does not propose new classification algorithms to improve the state-of-the-art biometric system performances. Instead, the contribution is to

present statistical analysis on the proposed architecture that demonstrates improvement in performance even when the individual classifiers used may not reach state-of-the-art error rates by themselves. A review of the multibiometric systems indicates three major issues in the design of multibiometrics that requires further research [31]. These issues are addressed in the dissertation for the proposed fusion scheme.

1.6.1 Classifier Fusion Architecture

In statistical pattern classification, the most notable progress has been made from the point of view of multibiometric architecture. The selection of appropriate system architecture depends on the application specifications and its accuracy, in general, improves with the increase in biometric information available for verification. The design of architecture first selects the biometric sources and then develops a methodology for identity verification. This design can then be extended to other sources by using some configurable rules to manage how each biometric source operates thereby allowing the user to improve the accuracy of the system dynamically.

The architecture designed for one application may or may not be considered optimal to a different application scenario. The application is tested by cross validating the design with biometric data set of the new application. However, this does not guarantee that the design is necessarily good but also the cross validation is costly in terms of the amount of data that has to be available. The architecture proposed in the dissertation can be generalised for multibiometric system with flexibility to choose the type of biometric, number and sources of biometrics [23]. The proposed architecture is experimentally evaluated for text dependent speaker verification using Hidden Markov Model (HMM) based digit dependent speaker models.

Analytical expressions for verification errors - false accepts and false rejects - are derived for sequential decision fusion of multibiometrics under independence assumption [32]. The main contribution is the empirical demonstration of the proposed architecture that effectively controls the verification error rates with potential improvement in performance even for weaker classifiers. The error trade-off is controlled by tuning the parameters related to the number of biometric sources. The architecture is also demonstrated to achieve superior performance despite the seemingly ideal assumption that classifiers make uncorrelated decisions [32].

Multiple instances and multiple samples are used as biometrics for the evaluation of proposed architecture. In the context of text-dependent speaker verification, multi-instance fusion concerns to the combination of information from different words or phrases spoken by an individual whereas multiple samples refer to the repeated utterances of a word/phrase spoken by the same individual. The dissertation also presents a theoretical and experimental analysis of multi-sample fusion which is novel in its treatment of random and adaptive multiple presentations within a sequential decision fusion architecture [33].

1.6.2 Classifier Correlation

In the context of multibiometrics, different biometric characteristics of an individual (e.g., face and speech) tend to be independent; however, for biometrics from a single modality (e.g., multiple speech samples of an individual) the independence assumption may not be true. Further, in multi-algorithmic fusion, the same biometric sample is used for classification by multiple classifiers and so the outputs can be expected to be highly dependent (correlated). Karthik et al. [34] have shown that for likelihood ratio-based fusion approach, the assumption of independence between classifiers does not adversely affect the fusion performance, especially when the individual's classifiers are accurate and the difference between genuine and impostor correlation is not high. Kuncheva et al. [35] have also shown that correlation in a classifier fusion ensemble can be both good or bad, and there are situations where the higher performance fusion can be achieved by considering statistical dependence between the classifiers.

The exact class-conditional error rates for the fusion of correlated decisions are estimated using full expansion of Bahadur-Lazarsfeld Expansion (BLE). The expressions for the error rates of the multi-instance fusion and multi-sample fusion schemes are modified to incorporate the correlation between the classifier decisions. The error rates for multi-instance fusion are developed considering the conditions of acceptance from each of '*n*' decisions. Similarly, the multi-sample fusion error rates are expressed using BLE and the conditions of rejections from multiple samples. The expressions for multi-instance and multi-sample fusion schemes are integrated for determining the proposed fusion verification error rates, i.e., the error rates for fusion of multiple samples were substituted as base errors in the expressions for multi-instance fusion error rates [36].

For statistically dependent classifier decisions, the error rates after decision fusion were higher (unfavourable dependence) or lower (favourable dependence) than when the

classifier decisions are *statistically independent*. Venkataramani et al. analysed the conditional dependence ('favourable/unfavourable') for the *'AND Rule'* [37] and *'OR Rule'* [38], using Q values between pairs of classifiers. However, the analysis for determining the statistical dependence between *'n'* classifiers is not fully explored [39]. This dissertation presents the expressions developed for determining favourable dependence between decisions from multi-instance and multi-sample fusion schemes that employ *'AND* and *OR'* rules. The developed expressions for verification error rates and conditions for favourable dependence are experimentally evaluated by considering the proposed architecture for text-dependent speaker verification using HMM based digit dependent speaker models [40].

1.6.3 Classifier Selection

In a multibiometric system, the amount of discriminatory information provided by each biometric source is quite different. Therefore, the classifiers modelled using these sources also differ in their ability to discriminate between client and impostors. It is, therefore, significant to select a set of classifiers that can be used to obtain optimal fusion performance. Given a fixed set of classifiers, there exist several criteria to determine which subset of classifiers achieves the optimal combination performance. The applicability of the *individual classifier performance* criterion for selection of classifiers is very limited because adding more and more poorer classifiers could only produce worse combinations [41]. Although diversity measures have been widely used for classifier selection, the experimental evidence [42] have shown very weak correlation between diversity measures and combination method performance. When the *combination method performance* is itself used for classifier selection, optimal performance is achieved at each stage of fusion. Therefore, consistent comparisons of different classifier subsets are obtained irrespective of the number of classifiers and their individual performances [43]. However, this method of selection with exhaustive evaluations can exponentially increase the complexity of the selection algorithm for large classifier pool [43].

The performance of a sequential combination method depends on the number of classifiers and the order in which the selected classifiers are fused for optimal performance. To avoid disagreement between the design of the combination method and the selection approach, the selection criteria should be tuned to the characteristics of combination method design. For sequential fusion of instances using *'AND Rule'*, lower total error rates (TER) are obtained when the *increase in false rejects* is less than the *decrease in false accepts*. Based on

these characteristics, the sequential error ratio (*SER)* measure is proposed that is the ratio of the number of samples for which classifier disagrees rather than agrees with previous classifiers correct decisions [44].

The SER based classifier sequence is used for better prediction of fusion performance on test dataset (unknown data) given the base classifier error rates and the variance in correlation coefficients from the tune dataset (known data) are also known. As the base classifiers for both the tune and test datasets were assumed to be similar in performance, the parameters (*n, m*) used to control the trade-off between FRR and FAR on tune dataset are also applicable to test dataset. Though the contributions are evaluated for speech modality, the framework can be applied to handwriting, fingerprint, keystroke dynamics and other modal characteristics.

1.7 Research Contributions

The research contributions of this PhD dissertation are as follows:

[1] V. Nallagatla and V. Chandran, "Sequential Decision Fusion for Controlled Detection Errors," in *13th International Conference on Information Fusion (FUSION)*, Edinburgh, 2010.

[2] V. Nallagatla and V. Chandran, "Sequential Fusion Using Correlated Decisions for Controlled Verification Errors," in *Computer Analysis of Images and Patterns*. vol. 6855, P. Real, et al., Eds., ed: Springer Berlin / Heidelberg, 2011, pp. 49-56.

[3] V. Nallagatla and V. Chandran, "Sequential Fusion of Decisions from Adaptive And Random Samples for Controlled Verification Errors," in *11th International Conference on Information Science, Signal Processing and their Applications*, Canada, 2012, pp. 793-798.

[4] V. Nallagatla and V. Chandran, "Sequential Fusion of Decisions from Multibiometrics for Controlled Verification Errors," in preparation for *Information Fusion, 2012*.

[5] V. Nallagatla and V. Chandran, "Sequential Fusion of Decisions with Favourable Dependence for Controlled Verification Errors," in *11th International Conference on Information Science, Signal Processing and their Applications*, Canada, 2012, pp. 259-264.

[6] V. Nallagatla and V. Chandran, "Classifier Selection using Sequential Error Ratio Criteria for Multi-Instance and Multi-Sample Fusion," in *6th International Conference on Signal Processing and Communication Systems*, Gold Coast, Australia, 2012, pp. 496-503.

Chapter 2
Information Fusion in Multibiometrics

2.1 Introduction

The concept of information fusion is studied under different terminologies such as classifier ensembles, hybrid methods, dynamic classifier selection, opinion pool or mixture of experts [1]. Ho [24] stated that the problem solving in pattern recognition has shifted from using the best features and the best classifiers to the best set of classifiers and then the best combination method. The goal of information fusion is to determine the best set of classifiers in a given problem domain and devise an appropriate function that can optimally combine the information provided by individual classifiers.

In the context of biometrics, information fusion refers to the use of multiple sources of biometric information to obtain a reliable decision [15]. Such systems, known as multibiometric systems, offer several advantages over the traditional single biometric systems. The consolidation of multiple biometric evidences results in an improvement of overall matching accuracy of the biometric system. The presence of multiple evidences/sources also effectively increases the dimensionality of the feature space and reduces the overlap between feature spaces of different individuals. These systems are more resistant to spoof attacks, as it is difficult to spoof multiple biometric sources simultaneously. Multibiometric systems also reduce the effect of noisy data due to the availability of multiple sources of information. This is especially important when verification has to take place in adverse conditions where certain biometric characteristics cannot be reliably extracted. For example, in the presence of ambient acoustic noise an individual's voice characteristics cannot be accurately measured.

From the literature in pattern recognition and information fusion, it is evident that the verification accuracy can be improved significantly by carefully fusing information in biometric systems. However, the issues related to the fusion architecture need to be considered before performing the fusion methods. Most of these issues depends on application requirements and impose few significant questions related to fusion design. In this chapter, the key issues such as different sources of biometric information (section 2.3), acquisition and processing architecture (section 2.4), level of fusion (section 2.5) and the

fusion methodology (section 2.6) for the proposed scheme are discussed in detail. The final section presents the architecture for multi-instance, multi-sample and the proposed fusion schemes with application scenarios.

2.2 Design Issues of a Multibiometric System

Multibiometric systems rely on the information presented by multiple biometric sources [1]. The primary design concern is the requirement for use of a suitable human-computer interface (HCI) that would permit the efficient acquisition of information from multiple sources with minimum inconvenience to the user. The major issues that are to be addressed in the design of the multibiometric system are:

- Sources of multiple biometric information
- Acquisition and processing architecture
- Type of fusion information
- Fusion methodology or combination method

The performance gain of a multibiometric system is measured as a function of the deployment cost. For cost-driven applications, the cost optimization aspect of the architecture is taken into consideration. With the increase in biometric sources/information used for fusion, the complexity of the verification system increases leading to higher cost, longer verification time and user inconvenience. When multiple modalities are combined, the user-interface has to be altered to acquire information from multiple sensors thereby increasing the cost of the system significantly. Therefore, biometrics from a single modality are generally preferred for reducing the system cost. For performance-driven applications with less consideration for the cost of the system, biometric sources employed for fusion could be from different well performing biometric systems. For this architecture, the system may require individual sensors for each modality, different feature extraction and modelling algorithms. Allano et al. [45] proposed a novel fusion strategy to reduce the cost of a multibiometric system by dynamically fusing the systems to optimise at the same time cost and performance of the system. The basic architecture of multibiometric system irrespective of performance or cost specifications should address the design issues that are discussed in detail in the following sub-sections.

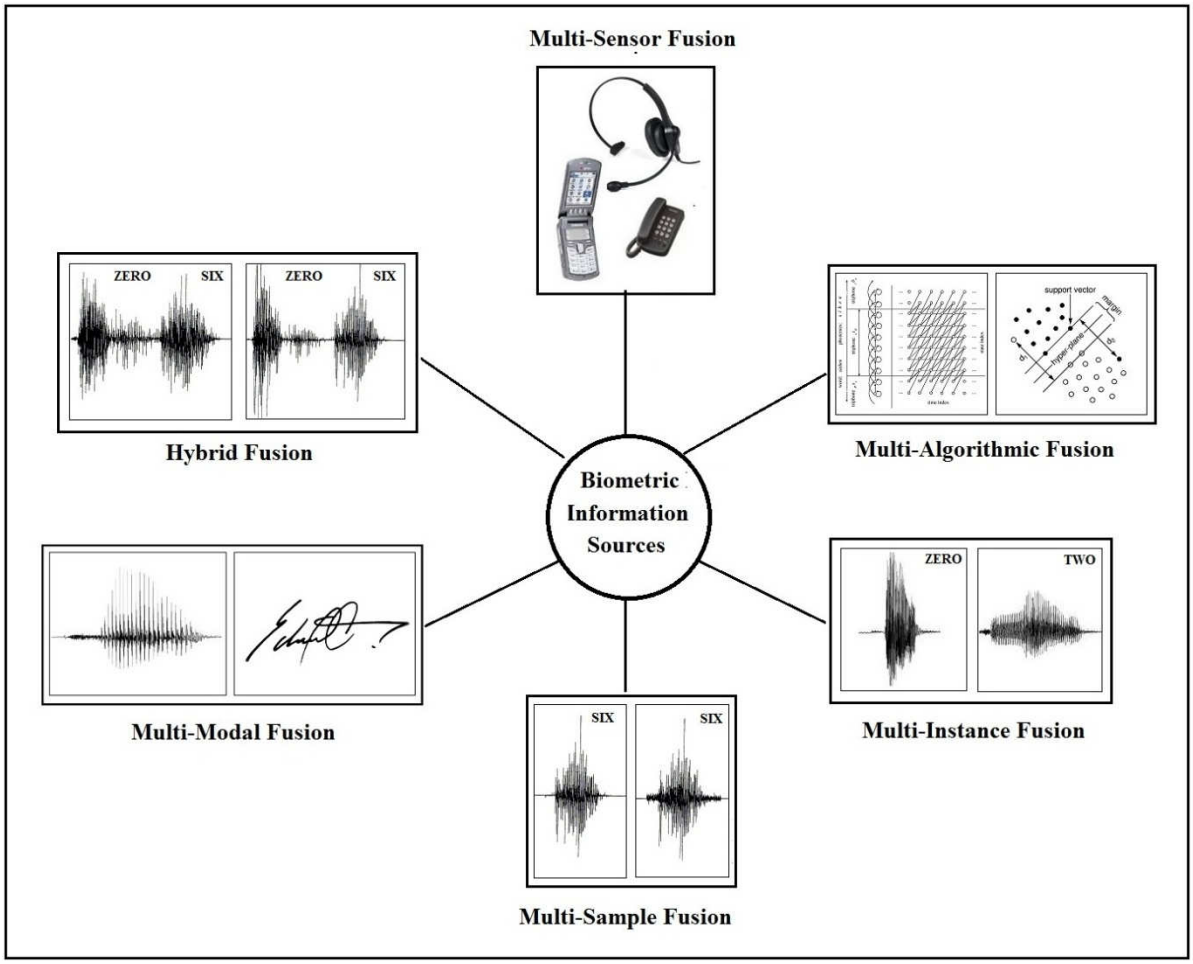

Figure 2.1 The various sources of information in a multibiometric system: multi-sensor, multi-algorithmic, multi-instance (waveforms for different verbal information), multi-sample (different waveforms for the same verbal information), multi-modal and hybrid fusion

2.2.1 Sources of Multiple Biometric Information

Multibiometric systems can be classified into six categories based on the nature of information source used for fusion [1]: multi-sensor, multi-algorithm, multi-instance, multi-sample, multi-modal and hybrid systems. In the first four scenarios, a single biometric characteristic is used for information fusion, while multiple characteristics are used in the fifth scenario. These scenarios (fig. 2.1) are explained below:

1. Multi-sensor systems: In these systems, biometric data from a single modality are acquired using different types of sensors. Lee et al. [46] proposed the use of multiple 2D cameras to acquire the face image of a subject for reliable face recognition. Multi-sensor systems are mostly used in the hope that diverse/complementary information

can be achieved from various sensors. The acquired data can be processed with one algorithm or combination of algorithms.

2. Multi-algorithm systems: Different processing and feature extraction methods for the same biometric data can result in different outcomes that may be desired source of information variation. The systems that employ various extraction algorithms may allow emphasizing different biometric features of interest (e.g., spectral or prosodic features of a voice sample) and produce different feature vectors for each. Multi-matching systems, on the other hand, allows matching the feature vectors against various types of models or/and matching techniques. These systems do not necessarily require the use of new sensors and are thus cost-effective compared to other types of multibiometric systems. Brunelli et al. [12] designed a multi-algorithmic and multi-modal system, i.e., two speaker recognition algorithms and three face recognition algorithms, combined at the match score and rank levels via a HyperBF network.

3. Multi-instance systems: The biometric data from the same biometric characteristic are acquired in terms of multiple instances or parts in a multi-instance system. These systems are also referred to as multi-unit systems. The design of the multi-instance system needs to determine the number of instances that are to be captured for a biometric. Multi-instance fusion for speaker recognition can consider the fusion of several samples of different verbal information (e.g., one, two, etc.) [47]. Similarly, the left and right index fingers, or the left and right irises of an individual, may be used to verify an individual's identity [48, 49]. However, systems capturing, for example, sequential frames of facial or iris images are considered multi-presentation rather than multi-instance. The cost associated with the multi-instance system is reduced when a single sensor is used to acquire the data in a sequential architecture.

4. Multi-sample systems: The multiple samples of a biometric characteristic are acquired to account for the intra-user variations of an individual or to obtain a more complete representation of the individual's characteristic. Cheung et al. [26] proposed the use of multi-sample fusion that corresponds to the combination of scores from multiple utterances of a speaker. In addition, a face recognition system, for example, may capture (and store) the frontal profile of a person's face along with the left and right profiles in order to account for variations in the facial pose. One of the key issues in a multi-sample system is determining the number of samples to be acquired from an individual. It is important to establish the desired relationship between the samples beforehand to optimize the benefits of the

integration strategy. For example, a face recognition system utilizing both the frontal- and side-profile images of an individual may stipulate that the side-profile image should be a three-quarter view of the face [50]. Alternately, given a set of biometric samples, the system should be able to automatically select the 'optimal' subset that would best represent the individual's variability [51].

5. **Multi-modal systems:** Multi-modal biometric systems utilise more than one physiological or behavioural characteristic for enrolment and verification. Multi-modal biometric systems can acquire input from single or multiple sensors measuring two or more different biometric characteristics. Physically uncorrelated characteristics [14] are expected to have better improvement in performance than the correlated characteristics [52]. The deployment cost of these systems is significantly high because of the change in the user interface to acquire information from different sensors. The number of modalities used for a specific application is limited by factors such as an increase in system cost, enrolment time, throughput time and expected error rate.

6. **Hybrid systems:** The systems that integrate a subset of above discussed fusion scenarios can be termed as a hybrid multibiometric system [53]. For example, Brunelli et al. [12] designed a multi-algorithmic and multi-modal system, that combines multiple speaker recognition and face recognition via a HyperBF network. A hybrid multi-modal system that combines a uni-modal face recognition and multi-algorithmic iris recognition systems is shown to improve performance [54]. Wang et al. [55] have shown that a multi-algorithmic and multi-instance fusion scheme for iris recognition can achieve better performance than uni-modal methods.

Another category of multibiometric systems combines primary biometric systems with soft biometrics (such as gender, height, weight, eye colour, etc.). Although soft biometric characteristics alone cannot distinguish individuals reliably, the biometrics when used in conjunction with primary biometric traits significantly enhance the performance of the verification system [56]. The hybrid multibiometric system that combines multi-instance and multi-sample fusion schemes with biometrics from a single biometric characteristic is employed for analytical and empirical analysis in this dissertation.

2.2.2 Acquisition and processing architecture

Once the sources of biometric information are known, it is required to decide the order in which the evidence will be acquired and then processed. The order or sequence of

biometric data acquisition has an impact on user convenience. Thus designing the acquisition protocol that enhances user convenience enables to reduce the enrolment time while ensuring good quality biometric data. In addition, the sequence in which the acquired biometrics is processed enables to determine the number of biometrics required for reliable verification. The acquisition and processing architectures can be either serial or parallel (fig. 2.2).

The acquisition sequence (serial or parallel) in a multibiometric system refers to the order in which the information from multiple biometric systems (where each system is provided with data from a single biometric source) is captured from an individual. It is usually convenient and cost-effective to acquire physically related biometric characteristics simultaneously (in parallel fig. 2.2 (b)). For example, face, voice and lip movement can be simultaneously acquired using a video camera [57]. On the other hand, sequential acquisition (fig. 2.2 (a)) is preferred when multiple instances of the same characteristic (e.g., iris images from both the eyes) or physically unrelated biometric characteristics (e.g., fingerprint and face) are considered for multibiometric verification.

Irrespective of the sequential or parallel nature of acquisition, the biometric systems can be processed in either serial or parallel mode to render a final decision. In sequential mode (figure 2.2 (a)), the biometric information from multiple sources is processed in a serial manner and so the decision is made before going through all the biometric systems. In the parallel mode (fig. 2.2 (b)), the processing of each biometric source is performed

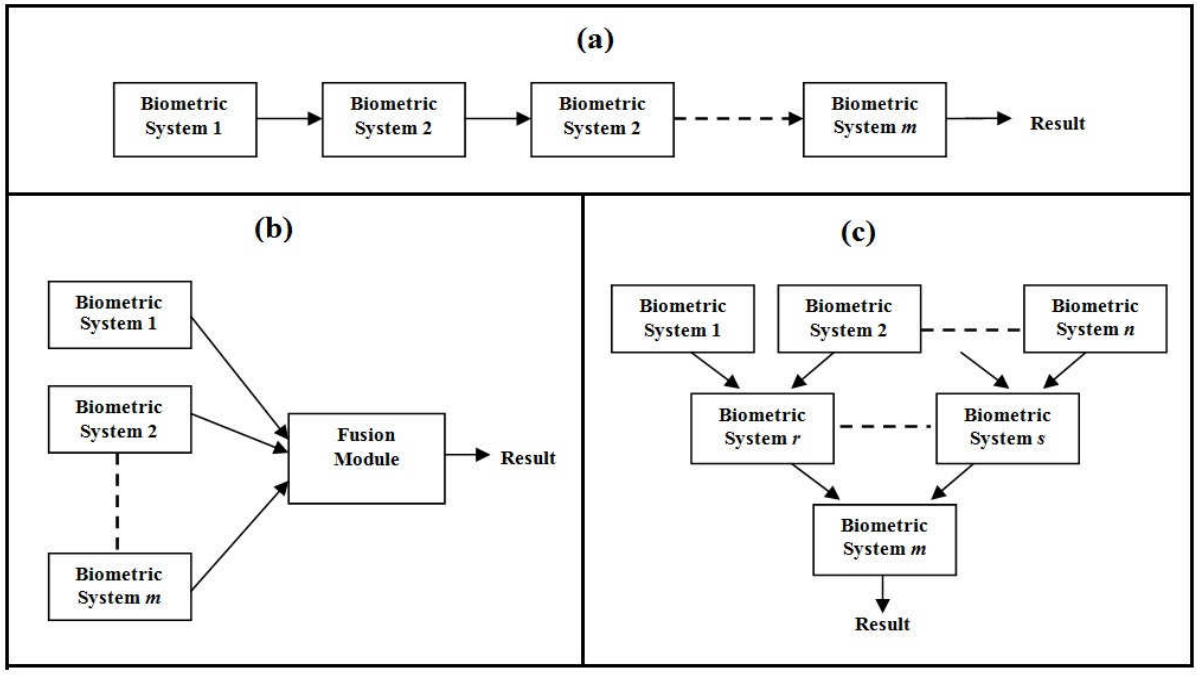

Figure 2.2 Multibiometric system architecture (a) Serial, (b) Parallel and (c) Hierarchical

independently at the same time and the information is combined using an appropriate fusion scheme. The advantages of both these modes of processing are combined using a hierarchical (tree-like) architecture (fig. 2.2 (c)) [5]. This architecture is applicable in situations where the acquisition or processing sequence is dynamically determined based on the availability and quality of individual biometric samples.

The choice of the system architecture, serial or parallel, is based on the requirements of an application. A multibiometric approach, in general, increases the system invasiveness and requires higher user cooperation. In case of genuine users, the average verification time of parallel fusion is same as that is required for the slowest biometric - in terms of both user cooperation and matching time. Serial processing of multiple biometrics offers better trade-off as only one biometric is acquired and processed initially and the system requires further acquisition/processing if there is not enough evidence for classifying the user [1].

As parallel processing architecture increases accuracy (because of the use of more biometric information), the architecture is more commonly used in the fusion literature where the primary goal of system designers is to reduce the error rates rather than processing time of biometric systems. This architecture is more suited for high security applications such as access to military installations [2]. Whereas for low security applications that are user-friendly such as bank ATM, the serial architecture can be used. Here, the user is given the choice to decide which source of information to be processed first. Therefore, a sequential processing system can be more convenient to the user and generally requires a shorter verification time compared to its parallel architecture [1, 2]. Akhtar and Alfarid [13] studied the robustness of multi-modal biometric systems against spoof attacks in serial and parallel mode of fusion. It is empirically shown that these systems in both fusion modes are not intrinsically robust against spoof attacks. When all the matchers are spoofed, systems in serial fusion mode can be most robust whereas parallel fusion mode is shown to be better when only the best individual matcher is spoofed.

The ***serial processing architecture*** [45, 58] has been poorly investigated compared to parallel architecture. However, Marcialis et al.[58] proposed a serial scheme that allows a trade-off between performance and matching time. They also explained a simple mathematical model able to predict the performance of two serially combined matchers. As this mode of operation enables to determine the biometric system used for fusion dynamically, the serial or sequential architecture is employed for the proposed multibiometric

system in which the outcome at any decision stage of multi-instance (or multi-sample) biometric system depends on the output of the preceding biometric subsystem.

2.2.3 Levels of Fusion

A multibiometric system consolidates information from multiple biometric sources. It is significant to determine the type of information that is best suitable for fusion of multiple biometrics. The fusion is performed at various levels and is categorised based on the processing modules used by a biometric verification system - sensor, feature extraction, matching, and decision modules [59]. The amount of information available for fusion decreases from the sensor module to the decision module (fig. 2.3). The raw biometric data (e.g., speech signal in the case of voice biometric) has the highest information content that is reduced by subsequent processing (e.g., after extraction of MFCC and generation of verification scores). In the verification mode, the final decision label contains only a single bit of information (accept or reject). Figure 2.3 shows the fusion at the various levels in a biometric system, and each source is introduced below.

- **Sensor level fusion:** The raw biometric data from the sensor(s) are combined in sensor level fusion [60]. Fusion at this level is performed only if samples from compatible sensors for the same biometric modality are combined or multiple instances of the same

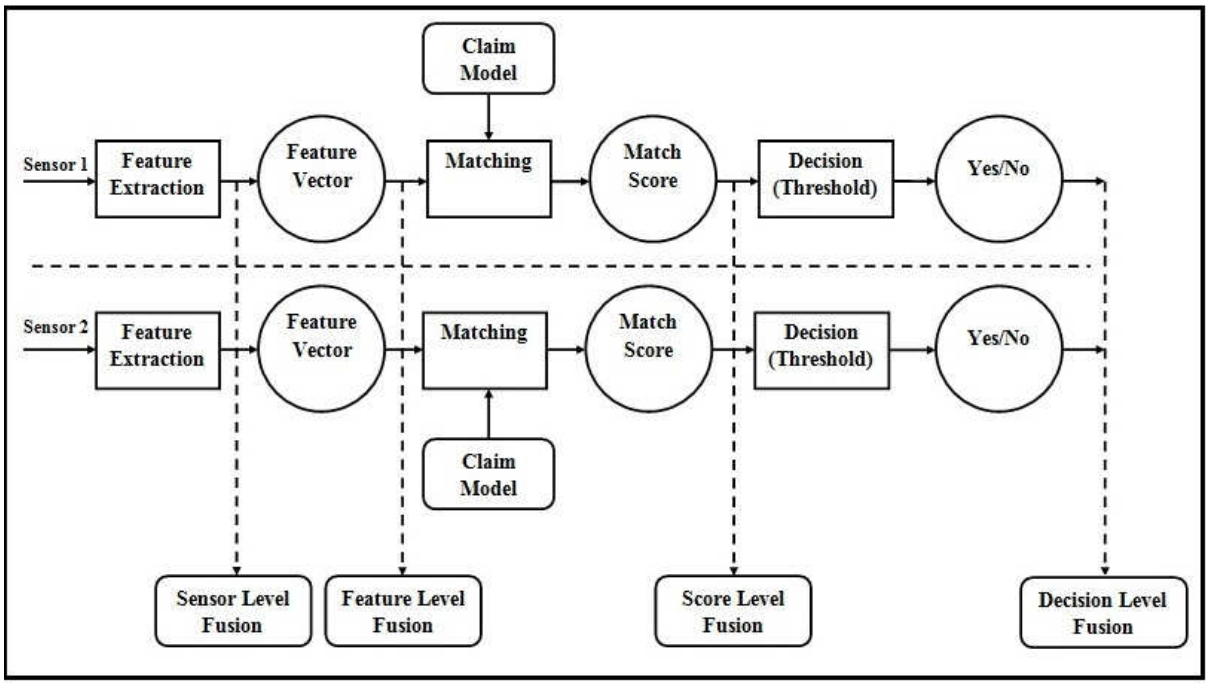

Figure 2.3 Levels of Fusion in Biometric System

biometric obtained using a single sensor are used for fusion. The multiple samples of an individual collected using multiple sensors enables to account for the intra-user variations and thus enhance the reliability of the data. The data must be combined in some meaningful way in order to enhance the subsequent verification accuracy of the system. For example, multiple 2D face images obtained from different viewpoints can be fused together to form a 3D model of the face [61].

- **Feature level fusion:** In this fusion level, the general approach is to extract and combine compatible features to form a large feature space for verification. The information thus obtained enables to complete the data profile for subsequent modelling and classification modules. The sensor readings are transformed into feature set by the application of appropriate feature normalization, transformation, and reduction schemes. Jain et al. [59] explained that it is important to ensure the feature attributes are as uncorrelated as possible at this stage to avoid biasing the classifier.

The problem with fusion at feature-level is that the feature spaces from different biometric sources may not be known or incompatible. If the two feature sets are fixed length feature vectors, then one could consider concatenating them to generate a new feature set. However, concatenating two feature vectors might lead to the curse-of-dimensionality problem [62] where increasing the number of features might actually degrade the system performance especially in the presence of small number of training samples.

- **Score Level Fusion:** In score-level fusion, the match scores for multiple biometric scores are combined to generate a new match score (a scalar) that can then be compared with a threshold at decision module for identity verification. Fusion at this level is the most commonly discussed approach in the biometric literature because of the ease of accessing and processing match scores (compared to the feature set extracted from the data). Jourlin et al. [63] have proposed an acoustic–labial speaker verification method where the scores from the visual features of a lip tracker and text-dependent features of speech classifier are combined using weighted sum.

The match scores generated by individual matchers may not be homogeneous. For example, one matcher may output a distance - a smaller distance indicates a better match, while another may output a similarity/matching score - a larger similarity value indicates a better match. Furthermore, the scores of the individual matchers may be correlated with different numerical scale (range) making match score level fusion a challenging problem.

- **Decision Level Fusion:** In decision level fusion, the independent decisions from different sources of biometrics are integrated. Methods like majority voting, weighted voting based on the Dempster-Shafer theory of evidence, *AND/OR* rules etc. are used to combine the decisions from biometric individual systems. In scenarios, such as commercial biometric systems, the systems provide access to only the final decision output (accept or reject) rather than scores. In this situation, the decision outputs that are combined do not include the confidence information or information on the strength of the match (strong or weak). Apart from biometrics, the decision level fusion has been widely applied in a number of areas such as multi-sensor data fusion [64], multi-spectral image fusion and geoscience data fusion [65, 66]. In some cases of multibiometrics, the term 'symbol level fusion' [67] is also used to represent decision level fusion.

Score-level fusion has been popularly used as scores are easy to compare and has rich information about biometric input. A disadvantage of score-level fusion is that the relationship between different scores may not be linear and it is difficult to construct a single similarity/dissimilarity metric for the combined score with considerable flexibilities. For example, different normalization methods of the matching scores lead to different decision boundaries. In addition, for methods with flexible boundaries the use of a too small training set of scores might easily overfit the data. In addition, the modelling of correlation between scores is shown to be useful only when the classifiers are of low accuracy and the difference between genuine and impostor correlation is large [68]. Further, the complete likelihood ratio based fusion rule, used for decision estimation, is based on the joint density of the genuine and impostor distributions and hence takes into account the correlation between the classifiers. The appropriate estimation of probability density distributions for more than two classifiers requires the score correlations of higher than second order.

The use of decision-level fusion framework is simple and clear from a mathematical point of view. Only a compact set of operation points is involved and the likelihood ratio criterion for decision-making is very beneficial for any biometric system. Furthermore, the use of decision fusion rules is very suitable for many real world biometric applications, with outliers existent in the genuine class [69]. Therefore, when the distributions of the genuine and impostor class are not symmetric, as is often true, the *'AND* and *OR'* decision fusion is very likely to fit because of their unsymmetrical support for genuine and impostor classes.

The probability distributions for fusion of multiple classifiers are estimated using equations for errors based on decision fusion. These equations may not be as straightforward

with score fusion as the final error will be a function of the threshold in joint probability density space and would be a multi-dimensional integral in general. The Bahadur-Lazarsfeld Expansion (BLE) is an approach used to approximate the probability density function for the statistically dependent case. The expansion applies to the simplified case of binary vectors (1-accept and 0-reject) for decision fusion and is simpler for decisions rather than scores. If score correlations of higher than second-order are to be considered in an analysis similar to the BLE, the analysis delves into the domain of higher order statistics, which is out of the scope of this dissertation.

The design of multibiometric system proposed in this dissertation combines information from multiple sources at ***decision level,*** not only because of the simplicity, but also of the possibility to build up a general fusion framework, without taking into account the specific type of biometric data processing and classification methods. The next section thus deals with the fusion schemes that can be employed to combine the decisions from multiple biometric sources.

2.2.4 Fusion Methodology

Decision-level fusion falls under a broader area known as distributed detection systems [70] and is defined as the process of selecting one hypothesis from multiple '*m*' hypotheses given the decisions of multiple '*n*' sources/classifiers. In biometrics, decision level fusion creates a single decision from typically two hypotheses (imposter or genuine user) of multiple biometric decisions. Decision-level fusion methods are often implemented to save communication bandwidth and improve decision accuracy. For a multibiometric system, fusion at decision level is described as the combination of decisions from a number of biometric systems that make verification decision independently for each source.

Several decision-level fusion methods and rules have been developed for biometric recognition over the past few years [71]. For example, Prabhakar and Jain [48] combined classifier selection and decision level fusion techniques to perform fingerprint verification. Kittler et al. [16] described a theoretical framework to derive a number of real rules for combining classifiers. The fusion rules, in general, can be subdivided into two main categories [72]: fixed rules such as *AND Rule*, *OR Rule*, majority voting, sum rule, and trained rules such as the weighted averaging of classifiers outputs, the behaviour knowledge space method. The theoretical and experimental results [73, 74] have shown that fixed rules usually perform well for ensembles of classifiers exhibiting similar performance. The

trainable rules generally are more effective for classifier ensembles exhibiting different accuracy or different pair-wise correlation [74]. In general, the individual biometric systems of multibiometric and multi-modal systems often exhibit significantly similar and different performance respectively [75]. Therefore, fusion of different biometric modalities is a task for which trained rules should perform better than fixed rules. However, the conditions of performance imbalance under which trained rules can significantly outperform fixed rules are not completely clear [74]. Moreover, in real applications like multi-modal biometrics, the bad quality and/or the limited size of training sets can quickly cancel the theoretical advantages of asymptotically optimal trained rules like the behaviour knowledge space [72].

The multiple biometrics employed in this work is from the same biometric characteristic - multiple instances and multiple samples. The fixed fusion rules that are employed for the proposed multibiometric architecture are briefly discussed below.

2.2.4.1 *AND Rule*

The principle of *'AND Rule'* for decision fusion is similar to that of Boolean *'AND'*. The identity claim of an individual is accepted for *'AND Rule'*, if only all the biometrics declare the claim to be authentic. The system with *'AND'* configuration provides high confidence that the individual who is introducing their biometrics to the system is who he claims to be. The use of *'AND'* rule makes it difficult to spoof multiple biometric sources thereby reducing the chances of accepting an impostor. In [76], the theoretical analysis of verification error rates (false accept and false reject rates) is explained for the combination of decisions using *'AND Rule'*.

Considering the false acceptance rate and false rejection rates of individual biometric subsystems to be α_i & ρ_i, $(i = 1, 2, 3, ...n)$ respectively, the resultant false acceptance rate (α_{AND}) and false rejection rate (ρ_{AND}) for the fusion of *'n'* biometric subsystems is given as

$$\alpha_{AND} = \alpha_1 \alpha_2 \alpha_3 .... \alpha_n \tag{2.1}$$

$$\rho_{AND} = \rho_1 + (1-\rho_1)\rho_2 + (1-\rho_1)(1-\rho_2)\rho_3 + ... + (1-\rho_1)(1-\rho_2)...(1-\rho_{n-1})\rho_n \tag{2.2}$$

The assumption here is that the decisions from individual biometric systems are statistically independent. The false acceptance rate for the combination of independent decisions is lower than that of any individual biometric system alone (2.1). However, the false rejection rate for the fusion of independent decisions is higher than any individual

biometric system alone (2.2). The false rejection rate for the fusion decreases if only one biometric system is used rather than combining systems with multiple biometric sources, especially if one subsystem is considerably stronger than other subsystems.

2.2.4.2 *OR Rule*

The principle of *'OR Rule'* for decision fusion is similar to that of Boolean *'OR*'. For the *'OR Rule'* to accept the user's claim, it is necessary for at least one of the biometric systems to declare the claim authentic. This *'OR*' configuration does not provide the confidence about the person identity claim as well as the '*AND'* configuration.

Considering the false acceptance rate and false rejection rates of individual biometric subsystems to be α_i & ρ_i, $(i = 1,2,3,...n)$, the resultant false acceptance rate (α_{OR}) and false rejection rate (ρ_{OR}) for the fusion of *'n'* biometric systems is given as

$$\rho_{OR} = \rho_1\rho_2\rho_3.....\rho_{n-1}\rho_n \tag{2.3}$$

$$\alpha_{OR} = \alpha_1 + (1-\alpha_1)\alpha_2 + (1-\alpha_1)(1-\alpha_2)\alpha_3 + ... + (1-\alpha_1)(1-\alpha_2)....(1-\alpha_{n-1})\alpha_n \tag{2.4}$$

The assumption here is that the decisions from individual biometric subsystems are statistically independent. The false rejection rate for the *'OR'* combination of independent decisions is lower than that of any individual biometric system alone (eq. 2.3). However, the false acceptance rate for the fusion of independent decisions is higher than any individual biometric system alone (eq. 2.4). When one biometric system in multibiometrics has a substantially higher EER compared to the other, the combination of the decisions using *'AND'* or *'OR'* rules may actually degrade the overall performance [63].

2.2.5 Other design issues

In addition to the above design issues, the performance of a multibiometric architecture depends on other factors such as:

a) Choice of Biometrics: The choice and number of biometric sources used for verification is largely dependent on the nature of the application, the overhead introduced by multiple biometric systems (e.g., computational demands and cost), and the correlation between the biometric information. Usually a choice-based combination of biometric sources produces better performance than arbitrary combination of information.

b) Multifactor Verification: Each type of verification systems (knowledge-based, possession-based or biometric-based) has its own strengths and weaknesses. The performance for the combination of different factors is improved when the architecture ensures that the strengths of one system complement the weaknesses of another. The design of the system needs a careful consideration on the type of knowledge or possession that is to be combined with biometrics for the user verification. For example, the user can present the biometric information of some specific knowledge, such as a password or PIN number, to the verification system. And the biometric characteristics extracted for verification can be either the user's uniqueness in uttering the knowledge information, his writing style or even the way the user types the information [23]. The design of multifactor verification systems should also consider user-inconvenience and other usability factors because a too inconvenient verification mechanism provides poor usability causing its users to revolt and find ways to circumvent it.

c) Trade-off between Cost and System Performance: The design of multibiometric systems with a high performance to cost ratio is an important goal. For multibiometric systems, high performance requirements usually necessitate information from many biometric sources, which results in an increase the system costs (including device costs). The cost here is a function of the number of sensors deployed, the time taken to acquire the biometric data, the storage requirements, the processing time of the algorithm and the perceived (in) convenience experienced by the user.

d) Catastrophic Fusion**:** The performance of a multibiometric system is supposed to be better than uni-modal biometric system due to the availability of additional information. And the phenomenon where the performance of a multibiometric system is below (one or more) individual biometric systems is known as catastrophic fusion [77]. For example, the audio-visual fusion approach at feature level has several disadvantages. The audio-visual feature vector has a larger dimension and due to the "curse of dimensionality" the parametric models, such as HMM's, trained on these features are less practical [78]. Also, feature fusion does not take the reliability of either modality into account; if one modality is very noisy, the audio-visual feature vector will be compromised and catastrophic fusion may occur; where the audiovisual accuracy is poorer than either of the single modalities [79]. Therefore, the design of multibiometric system should consider avoiding catastrophic fusion at all times.

e) User-specific parameters: Recent advances in multibiometric verification systems, especially those based on behavioural characteristics such as written signature [80] or voice

[81], have been accomplished by learning user-specific parameters. Examples of user-specific processing are user-specific feature extraction, user-specific model/template, user-specific fusion classifier [82], user-specific score normalization [83] and user-specific threshold [81]. User-specific processing is significant for applications where a fraction of users is more difficult to verify than the rest although the database is acquired in similar conditions. In practise, the reliability of user-specific statistics is greatly reduced with limited availability of user-specific data, especially for newly enrolled users. Toh et al. [84] have shown that at least six genuine samples are needed before its proposed user-specific procedure can outperform the baseline system. Such a large number of samples can be inconvenient in comparison to the conventional biometric applications that use only one sample. Therefore, it is significant to overcome the challenge to reduce the number of genuine training samples required for user-specific processing [85].

f) Trade-off between security and user convenience: The performance of a biometric verification system is mainly characterised by two verification error measures (*FRR* and *FAR*) that depend on the acceptance threshold '*t*' and the desired level of security. The *FAR* and *FRR* are related - *FRR* (t) is a monotonic decreasing function and *FAR* (t) is a monotonic increasing function. When the threshold setting is lowered to make it easier for clients, some unauthorized individuals may find it easier to gain access. The use of information from multiple sources with rigorous thresholds can make the system more accurate by improving the security against spoofing. Nevertheless, the acquisition of multiple sources causes user-inconvenience and takes longer time than the acquisition of single biometric source. As a result, the system design must consider appropriate decision threshold estimation methods to obtain better trade-off between security and user convenience (FAR vs. FRR).

In real-world applications, the ability to theoretically determine fusion performance using mathematical formulas instead of running experiments over again to test the new system could significantly reduce the required recourses. It is therefore desirable to derive closed form expressions for the fusion methods of a verification system. In addition, accurate error estimation information would be useful to configure appropriate thresholds and/or fusion rules will make the system more effective. The estimation of error rates for multibiometric system depends on the performances of individual biometric systems. In the biometric literature, the fusion performance is estimated for different ways in which the individual classifiers can be combined. Phillips et al. [10] provided the introduction to testing and error rates for general biometric systems. Golfarelli et al. [86] described a statistical

bayesian formulation of the errors (false accept and false reject rates) assuming the underlying distribution to be a mixture of normal distributions. Tumer and Ghosh [87] developed a theoretical framework for the analysis of simple average rule without the reject option. The authors also provided a theoretical and experimental analysis of the error-reject trade-off achievable by linearly combining the outputs of an ensemble of classifiers [88]. The practical comparison between different biometric combinations, when based on different technologies, is very hard to achieve. Therefore, analytical and theoretical results obtained under the assumption of unbiased and uncorrelated estimation errors along with simple guidelines for the design of multiple classifier (multibiometric) systems are feasible in theory.

The multibiometric system architecture proposed in this dissertation considers the combination of information from two biometric sources - instances and samples. The hybrid fusion scheme here reflects on the sequential integration of multi-instance and multi-sample fusion schemes at decision level using *'AND'* and *'OR'* Rules respectively. Analytical expressions of error rates are derived separately for multi-instance fusion and multi-sample fusion schemes for statistically independent classifier decisions. The proposed multibiometric architecture is then empirically evaluated to validate the developed expressions. The next section presents the architecture for the ***hybrid multibiometric system*** with detailed explanation and analytical analysis of multi-instance fusion and multi-sample fusion schemes.

2.3 Architecture of Hybrid Multibiometric system

The fusion of several biometric sources has been considered as a solution for the advancements in biometrics [15]. This combination of multiple sources can be divided into a loosely coupled solution and a tightly coupled solution. A loosely coupled solution assumes very little or no interaction among the inputs [11]. It integrates biometric data output of relatively independent sub-systems. An example of a loosely coupled system is the integration of audio and visual biometric data in an asynchronous manner. On the other hand, a tightly coupled solution assumes a strong interaction among the input measurements [89]. The biometric data is integrated at the sensor or representation level. An example of a tightly coupled system is the integration of audio and visual biometric data in a synchronous manner. The analysis of a multibiometric system here is performed using loosely coupled solutions where the biometric outputs are considered to be from relatively independent sub-systems.

A multibiometric system can be classified into one of the multi-sensor, multi-algorithm, multi-instance, multi-sample, or multi-modal systems based on the nature of the multiple sources [15]. Chang et al. [53] used the term ***hybrid*** to refer a systems that integrates a subset of the above mentioned classification of multibiometrics. Hybrid systems attempt to extract as much information as possible from the various biometric sources. For example, Brunelli and Falavigna [12] proposed an arrangement in which two speaker recognition algorithms are combined with three face recognition algorithms, i.e., combination of multiple algorithms and multiple characteristics, at the match score and rank levels via a HyperBF network. Poh et al. [90] evaluated an neural network classifier approach on a multi-modal system with face and voice biometrics. It has been shown that the use of multiple samples can boost the reliability of the multi-modal system. Similarly, a hybrid multi-modal system which is a combination of a uni-modal face recognition and multi-algorithmic iris recognition is shown to improve performance [54], i.e., the correct classification rate (CCR) of hybrid system gets increased to 99% where the CCR for face and iris recognition systems are 85% and 96% respectively. Wang et al. [55] have also shown that a multi-algorithmic and multi-instance fusion scheme for iris recognition can achieve better performance than uni-modal methods.

A multiple sample and multiple source approach is shown in [90] to increase the fault tolerance of system. Furthermore, this approach suggests that it is always beneficial to combine longer or more features (i.e., longer speech signal, more frames of facial features) to increase robustness without adding much cost to the existing system. Kevin et al. [91] presented the experimental results to indicate that multi-sample fusion techniques (*Multiple-Sample Single-Source*) can result in performance increase comparable to that of multi-modal (*Single-Sample Multiple-Source*) techniques. Multi-modal systems increase the abstract cost as additional hardware-software resources are required with more processing time for verification of each modality [45]. The cost of the system could be reduced by using different sources of information from the single modality, i.e., different instances or samples from the same biometric, or multiple algorithms or even the information from multiple sensors rather than multiple modalities [90]. Lorene et al. [45] proposed a novel *sequential fusion strategy* at score level to reduce the cost of a multibiometric system by dynamically fusing the optimal number of systems required to take the final decision. This method enables to optimise the cost and performance in the system simultaneously.

Sequential fusion [92], also named serial, cascaded or multi-stage fusion, has been used in literature for several different applications such as early verification decision [93]. In sequential approach, a test is done to find out if a decision can be made with some predefined degree of confidence. At each stage, if a decision is made with desired confidence the result of the test is accepted; otherwise, the decision is postponed until the next step where more data becomes available. This process is repeated until a final decision is made.

The sequential probability ratio test (SPRT) [94] is the most commonly used sequential fusion strategy that is based on the Neyman–Pearson theorem. Using this approach, desired biometric error rates can be pre-fixed and the testing is stopped at any time once a decision is made with enough confidence to support or reject the hypothesis [95]. Lorene et al. [45] considered this method to sequentially combine uni-modal systems in order to obtain a reliable decision with as fewer systems as possible with the aim of reducing the induced cost in a multibiometric system. Viola and Jones [96] also proposed a method in which a number of strong classifiers (built with standard AdaBoost algorithm in a sequence) are combined to form a cascade of classifiers of increasing complexity. Every stage of the cascade either rejects or passes the window to the next stage but only the last stage has the ability to accept the window. Thus, the window can be rejected at any stage but is accepted only after it passes through the whole cascade.

An approach that combines the benefits of these two schemes can be developed for biometric verification. A set of classifiers can be combined in a ***serial mode*** in which input is rejected at any stage but only accepted at the last stage, i.e., if accepted by all classifiers in the sequence. However, at each classifier level a ***sequential fusion*** approach can be employed where the testing can be repeated until the sample is accepted or testing reaches the (fixed) number of available data samples. These two methods are combined such that the strengths of one approach complement the weaknesses of another. This architecture is applied for a multibiometric scheme that combines information from multiple biometric sources of the same modality, i.e. ***multiple instances and multiple samples***. The design of this hybrid scheme considers the ***sequential/serial architectures*** for acquisition and processing of information from multiple sources that are considered independent. The decisions from these sources are combined (***decision level fusion***) to obtain a reliable final decision about the identity claim. The rule-based methods, ***'AND Rule'*** and ***'OR Rule'***, are employed for decision fusion in this work.

2.3.1 Multi-Instance Fusion Architecture

Multi-instance fusion refers to the combination of multiple instances of the same body characteristic for biometric verification. These systems also referred to as multi-unit system, have received considerable attention in the biometric literature. Examples are the use of left and right index fingers or the left and right iris of an individual to verify his/her identity [97, 98]. In such systems, where the data acquisition for instances can be sequential, in general, do not necessitate the introduction of new sensors whereas different sensors are required for simultaneous acquisition of different instances/units from an individual (section 2.4). Further, multi-instance systems do not entail the development of new feature extraction and matching algorithms and are therefore cost-effective.

Multi-instance systems are especially beneficial to users whose biometric characteristic cannot be reliably captured due to inherent problems. For example, a single finger may not be a sufficient discriminator for a person having dry skin but the integration of evidence across multiple fingers may serve as a good discriminator in this case. Similarly, an iris system may not be able to image significant portions of a person's iris due to drooping eyelids. The consideration of both the irises will result in the availability of more texture information that can be used to establish the individual's identity in a more reliable manner. Ramli et al. [47] studied the fusion of information from multiple instances of three verbal models (zero, seven and eight) using sum-rule and weighted sum-rule fusion. Jain et al. [99] have demonstrated that improvements are possible by combining two fingerprints or two versions of one finger for the test set evaluations of 160 persons. Jang et al. [98] also employed multiple instances, the left and right iris, as sources of information for authentication. A 2D dynamic programming-based minutiae matching algorithm [48] is adopted to get matching scores between the claimant and the enrolled multiple fingers where a Neyman–Pearson rule is used as the fusion scheme.

2.3.1.1 Applications Scenario

Biometric technologies provide user friendly and reliable access control methodology to computer systems, networks and workplaces [100]. Beattie et al. [101] discussed a scenario in which biometric sensors are placed at various locations in a building in order to impart security to individual facilities rooms. The authentication decision rendered at a particular zone (for a specific user) may depend on the decisions made previously in other zones (for the same user). The fusion scheme used to combine the decisions from multiple sensors can

also vary depending upon the zone sensitivity to verify the identity. For example, the *AND* decision rule may be used in high security areas - a user can enter such a zone only when all the sensors successfully confirm the individual's identity. Therefore, the scenario described above [101] permits the inclusion of multiple fusion rules involving multiple sensors in a dynamic architecture.

Similar architecture can be employed in most commercial biometrics applications (for example, telephone banking, access control or e-commerce) with consideration to *bi-factorial verification* (combination of knowledge and biometrics). The client/genuine user in these verification applications present biometric information of some specific knowledge (identification PIN/ credit card number/ password). The biometric characteristics extracted can be either the user's uniqueness in uttering the knowledge information, his writing style or even the way he types the information. For example, consider the scenario in which the user is asked to speak his account number. The identity claim in this case can be verified by classifying the entire account number at once (single instance) or by fusing classification information from individual characters (digit/alphabets) of the account number (*multiple instances*).

With a multi-instance system, each character is processed in sequence using a different classifier and so each instance has the ability to produce a decision about the user's claim independently. With knowledge-based authentication, user access is denied if at least one character in the account number is presented incorrectly. In bi-factor verification if one classifier (modelled for a character) in the sequence is rejected, the access claim to the system is denied. This fusion method effectively reduces the false acceptances, as it is hard for an impostor to reproduce a genuine user's characteristics for multiple instances. However, there is also a possibility that a genuine user is wrongly rejected at any stage of classification because of large intra-class variations. This method thus increases the number of false rejections. The approach of sequential instance combination is well suited for high security application scenarios, e.g., logging in as super user where providing access to unauthorized individuals is to be restricted to a minimum possible. The increase in FRR, however, causes greater customer inconvenience and so may not be desirable in most of the banking and point of service applications.

2.3.1.2 Framework of the Multi-instance biometric systems

The architecture of multi-instance system is shown in fig. 2.4. There is a sequential chain of classifiers $C_1, C_2, C_3, ..., C_n$ with each classifier for a biometric subsystem verifying an input test utterance $S_1, S_2, S_3, ..., S_n$ respectively. In this architecture, C_i refers to a classifier modelled for an instance '*i*'. From the scenario explained above, the user has to be accepted by all the classifiers (e.g., all characters in an account number) to allow the access. The *'AND Rule'* is used in decision fusion module as the final decision (*d*) of the system is to accept (*d*=1) the claim only if the decisions from individual classifiers $(d_1 = 1, d_2 = 1, d_3 = 1, ..., d_n = 1)$ is to accept the speaker.

The decision $d_i (i = 1, 2, 3.....n)$ for a classifier '*i*' is characterized by a false acceptance rate, α_i and false rejection rate, ρ_i. The decisions from each of the classifiers are assumed to be statistically independent and an *'AND Rule'* is used to fuse these decisions. From (2.1) & (2.2), the false acceptance rates and false rejection rates for the multi-instance *'AND fusion'* are shown to be lower and higher than that of any individual biometric instance respectively. When the error rates for individual biometric instances are considered similar, i.e., $\alpha_i = \alpha$, $\rho_i = \rho$ $(i = 1, 2, 3, ...n)$, the FRR and FAR for fusion of '*n*' instances is given as:

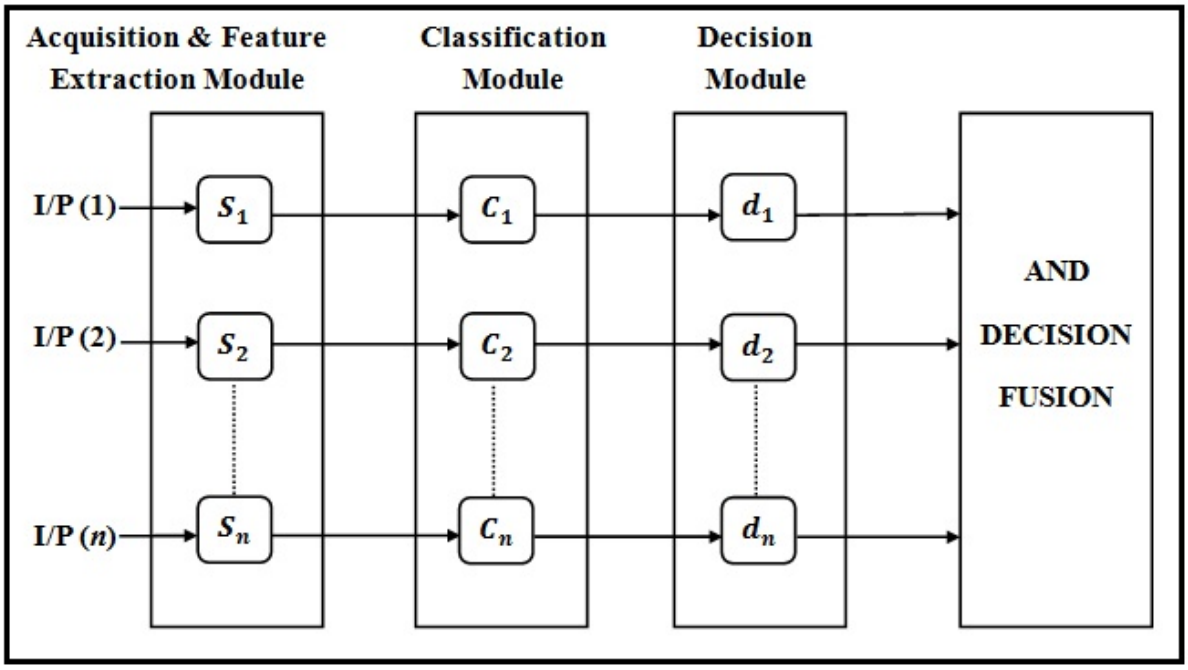

Figure 2.4 The architecture of multi-instance fusion of *'n'* instances

$$\alpha_{AND} = \alpha^n \tag{2.5}$$

$$= \rho\left(1+(1-\rho)+(1-\rho)^2+.....+(1-\rho)^{n-1}\right) = \left(1-(1-\rho)^n\right)$$

$$\rho_{AND} \approx n\rho \ \left(when \ \rho << 1\right) \tag{2.6}$$

The reduction in false acceptance rate is multiplicative (2.5) while the increase in the false rejection rate is approximately additive (2.6) for *'AND fusion'* which is desirable in most of the high security applications. The increase in false rejection rate can be reduced by using the complementary decision fusion *'OR rule'* for multi-sample fusion.

2.3.2 Multi-Sample Systems

Multiple samples of the same biometric characteristic can be acquired using a single sensor or multiple sensors. These samples can account for intra user-variations and/or to obtain a more complete representation of the underlying characteristic. For example, the information from multiple utterances can be combined for verifying a speaker [102], a face system can capture the frontal, left and right profile of a person's face in order to account for variations in the facial pose [103]. One of the key issues in a multi-sample system is determining the number of samples that have to be acquired from an individual. However, it is important to ensure that the captured samples represent the variability in individual's biometric data.

The desired relationship between the samples has to be established beforehand in order to optimize the benefits of integration strategy. For example, multi-sample fusion scheme is employed on a face recognition system that uses both the frontal and side profile images of an individual where the side-profile image is supposed to be a three-quarter view of the face [50, 104]. Alternately, Uludag et al. [51] explained schemes to automatically select the optimal subset of biometric samples that would best represent the individual's variability in the context of fingerprint recognition. The effectiveness of biometric system for the combination of decisions from multiple samples is based on decision-fusion techniques [102]. Although decision fusion is mainly applied to combine outputs of multiple modality-dependent systems, it can also be applied to fuse decisions from a single modality. The idea is to consider multiple samples extracted from a single modality as independent but coming from the same source [90]. This section investigates the *sequential 'OR' fusion* of decisions from multiple samples to improve the performance of biometric verification system.

2.3.2.1 Application Scenario

In a traditional password based systems, the user is allowed with certain number of attempts/tries (usually 3 attempts) to be verified by the system. Similar approach can be adopted for verification by replacing the password with repeated biometric samples from an individual. This method of multi-sample fusion helps in reducing the genuine user rejections but increases the false acceptances of an impostor. Similar to password-based verification systems, most biometric systems are designed to enable users to make more than one attempt for system access. While the statistical properties of successive password attempts do not necessarily relate to successive biometric-based verification attempts, an attempt limit is commonly suggested for both. Restricting the number of multiple samples to a minimum can limit the increase in FAR to certain extent. This is because, in practice, a false claimant/impostor usually requires more number of attempts to get accepted rather than a genuine user who will be good in adapting the biometric characteristics to his/her own model. Additional samples can be required by the genuine user for getting better acquainted (getting into position or adapting to the way of presenting biometric data, etc). This scheme of multi-sample fusion is mostly desirable in banking and online-transaction applications where a low false rejection rate is required for greater customer convenience.

2.3.2.2 Framework of Multi-Sample Biometric Systems

The architecture of multi-sample fusion system is shown in fig. 2.5. Here the classifier for the biometric system '*C*' is used to verify the input test samples $S_1, S_2, S_3, ..., S_m$ from the same instance. The maximum number of attempts allowed by the user in order to obtain access can be limited to '*m*'. For a user to be declared genuine, it is sufficient if any one sample presented to the system gets accepted and so an *'OR Rule'* can be used for acceptance. However, the user is denied access when all the '*m*' repeated samples are rejected and so *'AND Rule'* is used for rejection.

The decision $d_i \left(i = 1, 2, 3.....m \right)$ of a biometric system '*i*' is characterized by a false acceptance rate, α_i and false rejection rate, ρ_i. The decisions from each of these samples are assumed to be statistically independent. The false rejection rates for the multi-sample *'OR fusion'* is lower than that of any uni-biometric system (2.3). The true rejection rates for the multi-sample *'AND fusion'* is lower than that of any uni-biometric system (2.4). When the

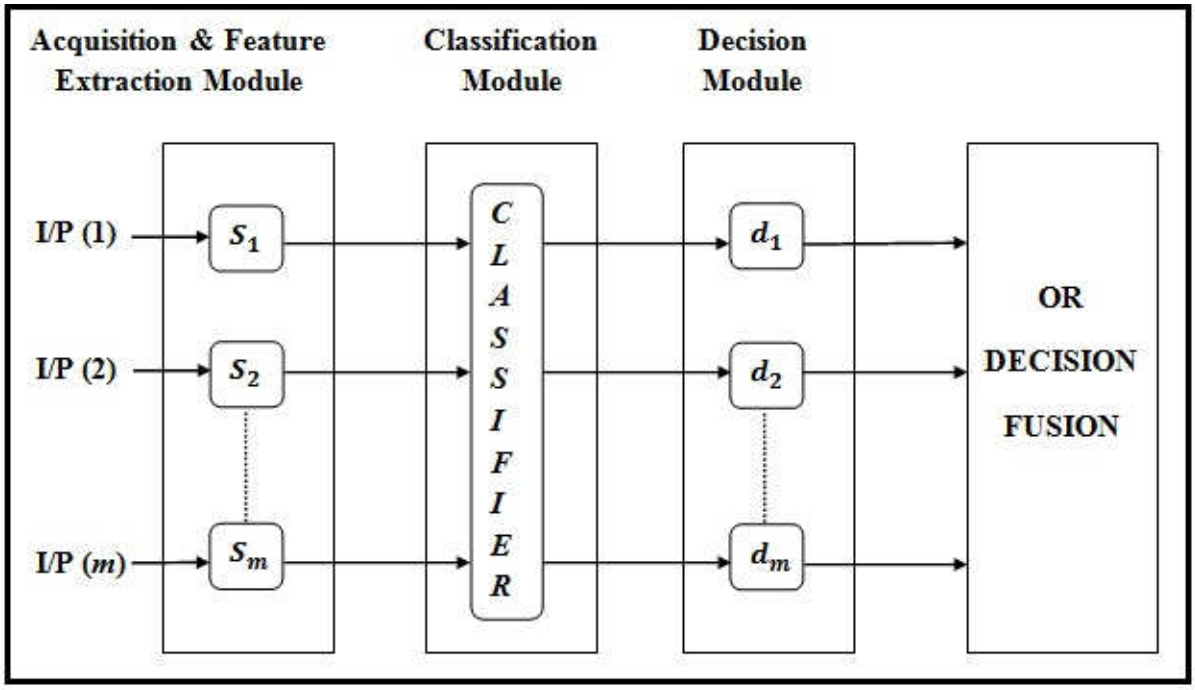

Figure 2.5 The architecture of multi-sample fusion of *'m'* samples

error rates for individual biometric systems are considered similar, $\alpha_i = \alpha,\ \rho_i = \rho$ $(i = 1,2,3,....m)$ for multiple samples, the FRR and FAR for fusion of '*m*' samples is given as:

$$\rho_{OR} = \rho^m \tag{2.7}$$

$$\alpha_{OR} = \alpha + (1-\alpha)\alpha + ... + (1-\alpha)(1-\alpha)...\alpha = \left(1-(1-\alpha)^m\right)$$

$$\alpha_{OR} \approx m\alpha \ \ (when\ \alpha << 1) \tag{2.8}$$

The reduction in the false rejection rate is multiplicative (2.7) while the increase in the false acceptance rate is approximately additive (2.8) and is desirable in most of the banking and point of service applications where user convenience is of importance.

2.3.3 Fusion of Multi-Instance and Multi-Sample fusion schemes

Biometric verification, in general, can be viewed as a serial process involving an acquisition (sensor), feature extraction, a classification and a decision module. Such a serial process could accumulate errors and reduce the overall reliability. However, the overall system reliability is increased by using several serial processes arranged in a parallel manner

[90]. The proposed hybrid approach employs the sequential *'AND fusion'* of multiple instances and sequential *'OR fusion'* of multiple samples.

The multi-instance and multi-sample fusion schemes reduce one type of error at the cost of increase in the other verification error. The integration of both these multibiometric schemes arbitrarily reduces both the verification errors [23]. Typical applications of the proposed architecture includes telephone and internet banking, information services, security control, remote access to computers, telephone and internet based shopping, etc. However, it is desirable in most of these applications to set the parameters, i.e., number of samples/attempts and the number of instances, used for verification of a specific speaker before performing real-world verification. The tuning of parameters enables the adjustment of application to be either highly sensitive (which is slower and may require repeated samples, but hard to breach) or less sensitive (which is still fast but may result in some false verifications).

In the proposed architecture (fig. 2.6), the maximum permissible number of repeated samples *'m'*, and the number of instances *'n'* are fixed prior. In this system, the user presents

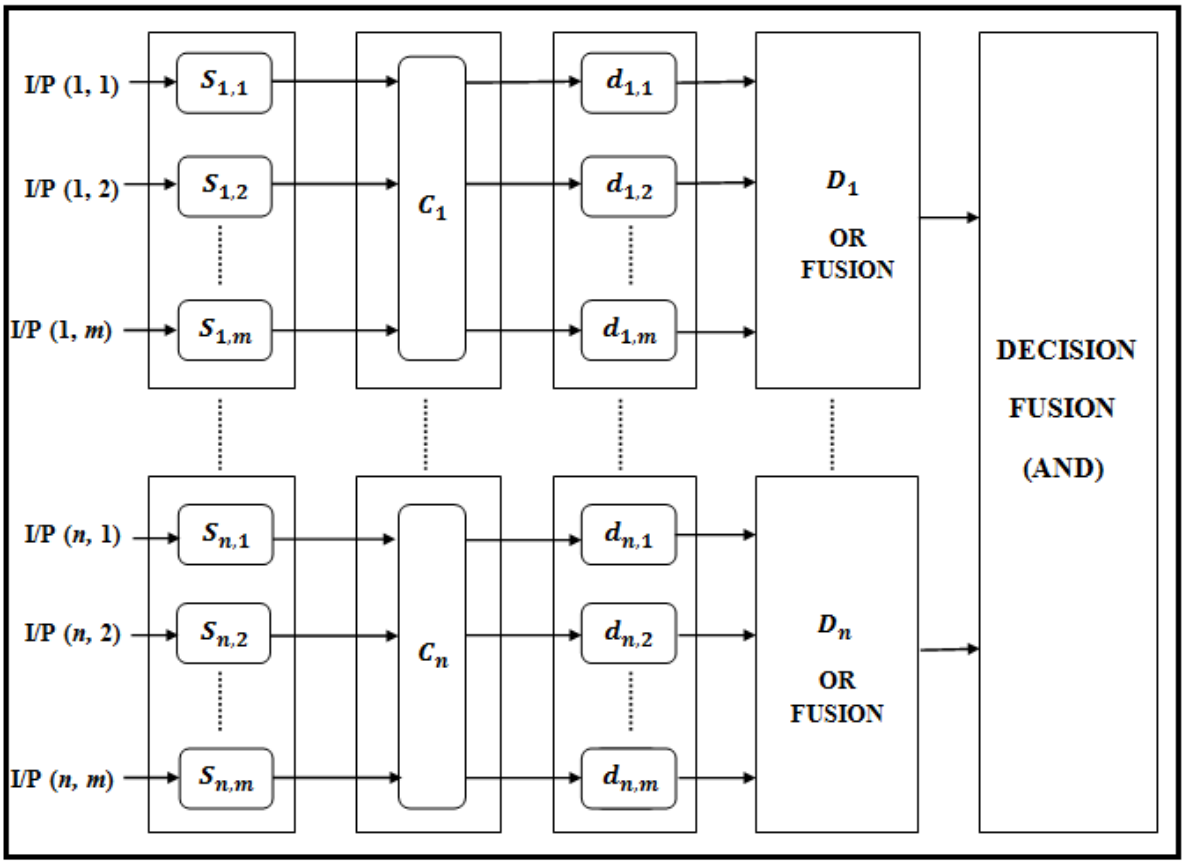

Figure 2.6 The architecture of multi-instance and multi-sample fusion schemes

an input test sample $S_{i,j}$*(i=1, 2 ...n, j =1, 2 ...m)* and the biometric system C_n makes a decision to either accept or reject the claimed identity. For a user to be declared genuine for a particular instance (e.g., spoken text), it is considered sufficient if any one sample presented to the system gets accepted. Acceptance decisions are logical *'OR'* for multiple samples. However if the user is accepted by 'i^{th} sample' ($1<i<m$) then the subsequent samples need not be verified. The user is considered to be an impostor when all the '*m*' samples are rejected. Rejection decisions are logical *'AND'* for multiple samples. Conversely, it is considered necessary in the sequential decision framework that a user be accepted by all instances in the sequence of decision stages. Acceptance is thus logical *'AND'* for multiple instances. If the user is rejected by any decision stage, the sequence terminates and thus rejection decisions are logical *'OR'* for multiple instances.

Considering false acceptance rate (FAR, α) and false rejection rate (FRR, ρ) to be independent for each instance, the fusion scheme expressions are given as:

$$\text{Multi-Sample Fusion}: \alpha_{OR} \approx m\alpha \ \left(\text{when } \alpha << 1\right); \rho_{OR} = \rho^m$$

$$\text{Multi-Instance Fusion}: \alpha_{AND} = \alpha^n; \rho_{AND} \approx n\rho \ \left(\text{when } \rho << 1\right)$$

$$\text{Multi-Instance \& Multi-Sample Fusion}: \alpha_{(n,m)} \approx \left(m\alpha\right)^n; \rho_{(n,m)} \approx n\rho^m$$

From the above equations it is clear that while the FRR decreases (since ρ is less than 1) multiplicatively with the number of attempts *'m'*, the FAR increases additively with *'m'* and the reduction in the FAR is multiplicative with the number of instances *'n'*, while the increase in the FRR is approximately additive with *'n'*. The facts to be noted here are (a) the behaviour with respect to *'m'* and *'n'* are complementary and (b) multiplicative changes are faster than additive ones and this enables the control of the errors through these parameters in the architecture.

With the above equations, it is possible to design a fused system that has lower errors of both types compared to a single verification stage using a single sample. It is also possible to keep both errors within reasonable bounds – without false rejections rising quickly to nearly 100% when the false acceptance reduces or the other way around. The trade-off in achieving this is the computational time required to perform multiple matches and make decisions with every sample and instance in the architecture. It will indeed be so if the

decisions were statistically independent as assumed, between multiple samples as well as for multiple instances.

The above architecture is empirically evaluated by verifying the identity claim of an individual using his/her unique speech characteristics. Text-dependent Hidden Markov Models (HMM) are created for different instances (e.g., word, phrase, etc) of a speaker. The training and modelling techniques used for text-dependent speaker verification are explained in next chapter. It also describes the challenges that are to be considered when designing the speaker verification system. The databases used for evaluation are also explained in detail.

2.4 Chapter Summary and Conclusions

In the context of biometrics, information fusion refers to the use of multiple sources of biometric information to obtain a reliable decision. Such systems, known as multibiometric systems, can improve the accuracy of a biometric system. When multiple biometrics are acquired and processed in a sequential fashion (cascaded combination) the time required for generating a decision reduces significantly. As the architecture for fusion at decision-level is simple and clear from a mathematical point of view, the independent decisions from multiple biometrics are combined for better accuracy.

Based on the nature of information sources being consolidated, multibiometric systems can be classified into six categories - multi-sensor, multi-algorithm, multi-instance, multi-sample, multi-modal and hybrid. The design of the system needs to consider the efficient acquisition of high quality biometric data. The task of choosing the information from biometric sources for fusion can affect the verification performance. Factors such as cost, throughput time, user convenience play a large role in selecting the biometric sources and adopting a particular fusion strategy. The other design issues of significance are multi-factor verification, user-specific processing and trade-off between security and user convenience.

The security and user convenience factors for an application are dependent on the verification threshold. The tighter threshold, in general, has the advantage of higher security (i.e., lower FAR) whereas lower threshold provides higher user convenience (i.e., lower FRR). An architecture based on the sequential integration of multiple instances and sequential fusion of multiple sample using *'AND and OR Rules'* is proposed. This method is theoretically shown to allow a controlled trade-off between false alarms and false rejects when the classifier decisions are statistically independent. Equations developed for

verification error rates are experimentally evaluated in the next chapter by considering the proposed architecture for text dependent speaker verification using HMM based digit dependent speaker models. The architecture investigated is applicable to speaker verification from spoken digit strings such as credit card numbers in telephone or VOIP or internet based applications.

Chapter 3
Text-Dependent Speaker Verification

3.1 Introduction

The sequential decision fusion architecture discussed in previous chapter is evaluated by performing verification on voice or speech patterns of an individual. The speech patterns, in general, are seen to be a mixture of physiological and behavioral characteristics of biometrics [4]. The physiological properties of speech/voice generally offer more intrinsic security whereas behavioral properties have the ability to adapt easily to the changes in the behaviour of the user. Therefore, speech is attractive for remote authentication approaches that require liveliness detection and anti-replay attack mechanisms to distinguish between an original and a fake biometrics.

Voice biometrics offer an advantage in that it is natural and easy to produce with the requirement of little custom hardware [6]. It can be captured non-intrusively and conveniently with simple transducers and recording devices. The task of validating a speaker's identity claim using these captured speech samples is known as Speaker Recognition [105]. Recognition using speech samples is highly accurate (in clean noise-free conditions) with low computation costs as the samples contain not only the message being spoken but also the information about the voice production system of the speaker [106]. The information extracted from these speech samples is distinctive among the speakers for the differences in voice production mechanisms such as the length of the vocal tract, characteristics of the vocal cord and the differences in their speaking habits [106].

Voice biometrics has a wide range of applications because of the pervasiveness of speech signals. The recognition in these applications can be performed considering the speaker to be either cooperative (e.g., to be given access to a specific system such as his/her bank account) or non-cooperative (e.g., confirming his/her presence at home in an automatic home parole control application). Speaker recognition is being currently used in conventional physical access control [107] and the broader area of remote identity recognition [108] as may be utilised in online transaction processing and interactive voice response [109] applications (such as banking over a telephone network, information and reservation services, telephone shopping, voice dialling [110] and voice mail). Speaker recognition also has

applications in Law Enforcement, i.e., Surveillance Applications and Forensic Investigation [111].

The task of real-time speaker recognition in everyday life is to improve security and are sometimes used in conjunction with other biometric recognition techniques like face, fingerprint [14] to improve security. Use of speech for principle security control is still not fully preferred because of drawback that speech samples are subject to some sources of variability that will reduce the recognition accuracy. The variability can be involuntary, for example, a speaker's inability to repeat samples precisely the same way, whereas some speakers can attempt to disguise their voices voluntarily. Other sources of variability can be because of variations in background noise, transmission and recording conditions. Several normalization [112, 113] and adaptation techniques [114] are proposed to minimise the effects of these variability on the performance of speaker verification. The other issues such as channel variations, impostor data for threshold selection, limited training and testing data need to be addressed while designing the verification system.

This chapter provides basic understanding of speaker verification technology with brief explanation on the classification of speaker recognition. The next section 3.3 deals with the basic architecture of the text-dependent speaker verification system along with the training, modelling and decision making techniques used for HMM based text-dependent speaker verification. Section 3.4 discusses the deployment issues to be addressed for the design of the text-dependent speaker verification architecture. The database and protocols used for the experimental analysis are discussed in section 3.5. This section also presents the baseline performances of text-dependent speaker verification for different data *SET*s.

3.2 Classification of Speaker Recognition

Speaker recognition is the task of recognizing people based on their voice characteristics [105]. The recognition task requires the extraction and modelling of acoustic features of speech that are unique to the speaker. This process of recognising speakers can be classified based on the task and text used for recognition.

3.2.1 Task Dependence

Based on the type of application, speaker recognition can be classified into two specific tasks [115]: Speaker Identification and Speaker Verification. In speaker

identification, an unknown speaker's voice/speech sample is compared with a set of known speaker models, and the best matching speaker is taken to be the identified speaker. This task can be referred as *closed-set identification* [115] when the set of speaker models include all speakers of interest. Most of speaker identification applications are *open-set*, in which it is possible that the unknown speaker is not included in the set of speaker models. If no satisfactory match is obtained for *open-set*, a *no-match* decision is provided. In speaker identification, the performance degrades with an increase in speaker models and comparisons.

In speaker verification [116], an identity claim is provided along with the speech sample. In this case, the unknown speech sample is compared only with the claimed speaker model. If the similarity is satisfactory, the identity claim is accepted, otherwise the claim is rejected. Speaker verification is considered as a special case of open-set speaker identification with a one-speaker target set. As speaker verification requires only one comparison, the performance of speaker verification system is independent of the size of speaker population.

3.2.2 Text dependence

Speaker recognition systems can also be classified into text-independent and text-dependent systems based on the dependencies on texts. The text-independent system enables the speaker to use any sentence/word of his own choice as the speech sample for identity recognition. The system may also prompt the user to provide with an unpredictable text as a sample for recognition (*user-driven text-independence*). The text-dependent system (*machine-driven text-independence*) usually requires the (phonetically) same text/speech sample (text includes phrases, words, syllables, phonemes, etc.) for training and testing the speaker.

The text-dependent systems can be further classified based on the type of text used for recognition [117]:

- ***Text-dependent using a fixed vocabulary shared by all users***: The same text is used during training and testing for all set of enrolled speakers [83]. Though this method could be used for test speaker discrimination, the system is not likely to be used in a real application.

- ***Text-dependent using fixed user-dependent vocabulary***: The speakers enrolled in the system can be trained on different text specific for each speaker. However, each speaker uses the same text during enrolment and test phases [118].
- ***Vocabulary-dependent***: The speakers are enrolled for all text/words in a given vocabulary (e.g. digits 0 through 9, spelling-word sequences, or a small set of arbitrary words [105]). The testing is performed on a subset of text from the vocabulary in either the same or the different order in which the words/phrases are enrolled.
- ***Event-dependent***: The speaker is recognised using models for particular events in the speech signal, e.g. particular phonemes, word or the occurrence of grammatical errors. The text used for enrolment and test phases can be either same or different as long as the modelled events occur in sufficient numbers [119].

The limitation with text-dependent and independent systems is that they can be easily deceived by the replay attacks where an impostor with access to the recorded sample of a registered speaker can be falsely accepted as the claimed speaker. Text-prompted (*text-dependent machine-driven*) systems [120] can be used to deal with this problem where the system requires the user to utter a specific text, in particular order, which is generated independently for each recognition. A small set of words, such as digits from *'zero'* to *'nine'*, can be used as key words and the speaker can be tested based on the generation of arbitrary sequence using these key words. The use of prompts makes replay-attacks more difficult, as the recording of a single-pass phrase is not sufficient to fool the system. The disadvantage, however, is that longer speech signals are needed for training and testing and so the system is not as convenient as the single pass phrase approach.

The choice of the text dependence for a classifier depends on the application requirements. The choice of classifier and configuration depend on certain application constraints such as [121]:

- The level of *user cooperation* determines the state of the system, i.e., an active or passive system. If the speaker is cooperative, the system can ensure the liveliness of the speaker by prompting for additional input speech. However, if the speakers are uncooperative the system is considered passive. For example, text-independent systems are used in forensic and surveillance applications where the user may not need to be cooperative and often not aware of the task. Whereas, text-dependent systems are mainly

used in access control and voice authentication applications where the speaker provides the input samples.

- High *verification accuracy* may be a requirement in most of the remote authentication areas such as access to a banking account where high accuracy is desirable in providing accurate access to a user's account. A text-dependent system is usually applicable in this case since it offers higher performance than text-independent techniques.
- The *amount of speech data available for training and recognition* also helps to determine the type of classifier. Nevertheless, obtaining enough training data is one of the biggest practical challenges in choosing a classifier in real-world applications. If more data is available for each text in the vocabulary, a text-dependent system can be employed otherwise a text-independent system can be trained from the limited combined data.

The other constraints include available computation and memory resources, channel usage or how the output is used. The performance of a text-dependent and/or a text-independent system is mainly dependent on the constraints of the application scenario. It is difficult to characterize the accuracy of speaker recognition systems exactly in all applications due to the complexities and differences in the enrolment and recognition scenarios. Lam [122] has shown that the range of performances for text-independent and text-dependent speaker verification systems with consideration to application dependent constraints. It has been mentioned that the system performance improves as more constraints are placed on the application scenario (e.g., increase in amount of training data, more benign channels).

The sequential architecture proposed in chapter 2 is applicable to the scenario where multiple instances and samples are combined to produce a reliable final decision about the identity claim. Multi-instance or multi-unit systems fuse multiple instances/parts of the same biometric characteristic for verification. In case of verification based on speaker's characteristics, each of the different ***verbal terms*** (such as different words or phrases) is referred to as an ***instance***. The decisions from the classifiers modelled separately for each word/phrase (i.e., text-dependent systems) can be considered independent and when combined can result in improved fusion performance. A multi-sample system combines multiple samples/tries that account for the variations in the same biometric. ***Multiple utterances*** for the same word/phrase from a speaker represent ***multiple samples (repeated samples)*** acquired for an instance.

The sequential architecture is theoretically analysed (in chapter 2) under the assumption of statistically independent classifier decisions from independent sources of information. However, this assumption is an ideal one and is difficult to achieve in practice. For a text-dependent speaker verification based on speaker-dependent HMM classifiers for each word, the assumption of independence is good when the phonemes involved in the word are different and will hold reasonably well even when they share some phonemes but differ in the order in which they are put together [23]. The text-dependent speaker verification system is thus employed for empirical evaluation of the proposed sequential decision fusion architecture. The next section explains the issues related to the design of the text-dependent speaker verification.

3.3 Architecture of Text-dependent Speaker Verification

Speaker verification is a one-to-one mapping between a speaker's voice and a claimed identity's voice. A speaker can provide the identity claim by entering a pin or using a smart card. A speaker verification system is composed of two distinct phases, a training phase and a verification phase each of which can be seen as a succession of independent modules. Figures 3.1 (a) and (b) illustrates a modular representation of training/enrolment phase and verification/test phase of a speaker verification system respectively. The first step in both the training and verification phases is the extraction of feature parameters from a speech signal (section 3.3.1). Feature parameters extracted are used in the speaker-modelling step to create a model for each of the speakers' voices (section 3.3.2). Feature parameters extracted from the test utterance are matched against the reference speaker model created during training phase. The output of this comparison is a similarity score that is used to decide whether to accept or reject the claimant (section 3.3.3). The text-dependent speaker verification system developed for this dissertation work utilises the ***HTK Toolkit*** [123] for the *feature extraction (MFCC)* and *HMM modelling techniques*. These modules are discussed in detail in the subsequent sections.

3.3.1 Feature Extraction

The recorded speech data is in the form of a continuous speech waveform. For efficient speaker verification, the speech waveform is normally converted into a sequence of time-discrete parametric vectors that can be referred to as feature vectors or observations.

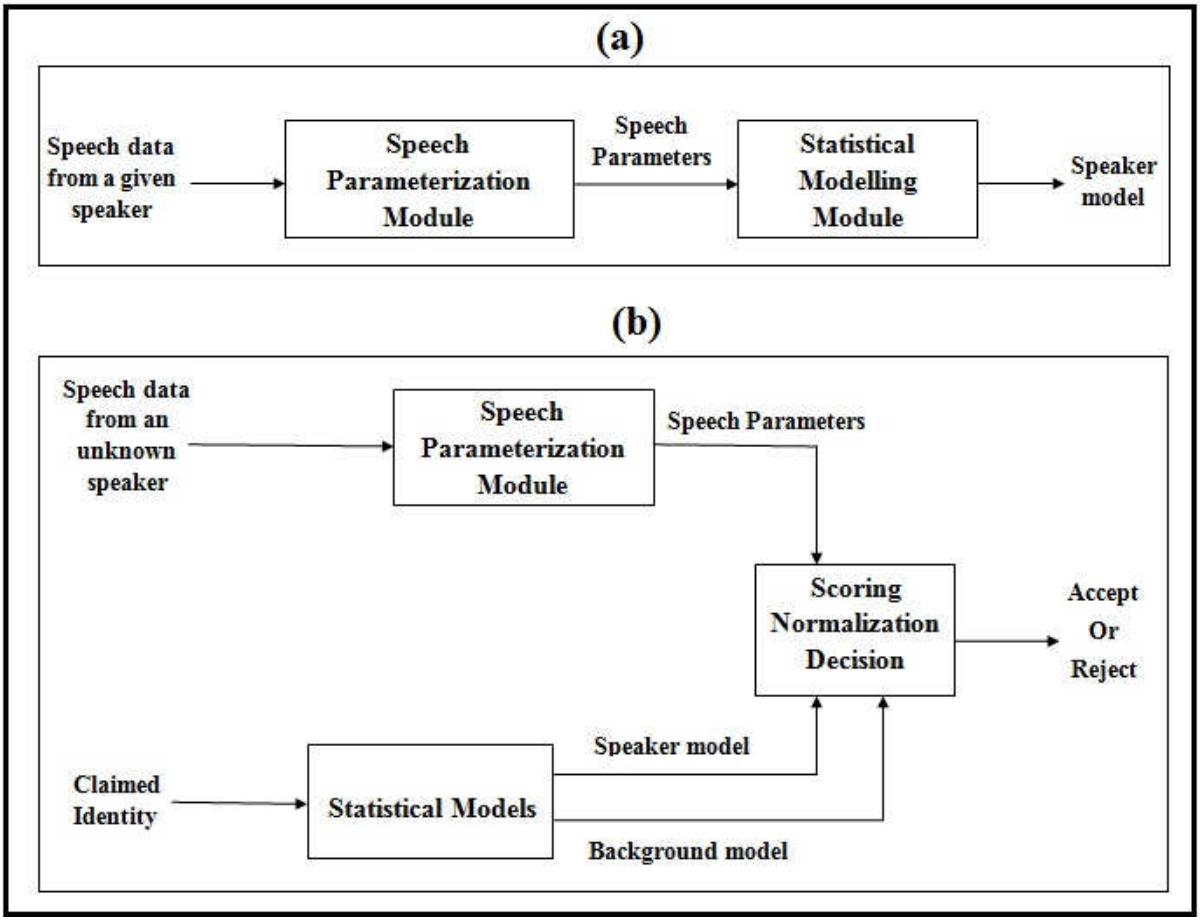

Figure 3.1 Training and verification architectures for speaker verification

This process is known as feature extraction and is assumed to give exact and compact representation of speech variability. The choice and number of types of features extracted influence the performance of the whole verification system. “The curse of dimensionality” problem arises when many different feature extraction methods are used [115]. The more features one uses, the larger the feature dimensions become and consequently, the increase in the complexity of computing. Hence, it is very important to understand the advantages and disadvantages of the different features and to use only the ones most relevant to the problem at hand.

The two main speech parameterization techniques used for speaker verification systems are Linear Prediction Coding [116, 124] and Filter-bank analysis [125]. The linear prediction method is based on a powerful speech production model quite suitable for voiced sounds and still acceptable for unvoiced sounds. The filterbanks have shown a better behaviour in the presence of noise [126]. The popular filterbank features include Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictive Cepstral Coefficient (LPCC).

Currently, the predominant choice of parameters are mel frequency cepstral coefficients (MFCC) [127]. The MFCCs have been shown to be more reliable and robust feature vectors than LPC coefficients. The reason is that human perception of frequency content of sounds does not follow a linear scale [128]**.** The *Mel frequency cepstral coefficients (MFCC)* are used in this dissertation work for extraction of speaker characteristics required for training and verification of a speaker.

First, the speech signal is split into discrete segments usually with 10ms shifting rate and 32ms window length. This reflects the short-term stationary property of speech signals [129]. These discrete segments are often referred to as *frames*. A feature vector is extracted for each frame. A pre-emphasising filter is normally used during the feature extraction to boost higher frequencies. The filter has the transfer function

$$x_p(t) = x(t) - a * x(t-1) \tag{3.1}$$

Where $x_p(t)$ is the pre-emphasized sample, $x(t)$ is a raw signal sample at time '*t*' and 'a' is the pre-emphasis coefficient which lies in the interval [0.95, 0.98] (coefficient - 0.97 in this work). The Hamming window used can be given as

$$w(n) = 0.54 - 0.46 * \cos\left(2\pi \frac{(n-1)}{L}\right) \tag{3.2}$$

Where *n* is an integer index $0 < n \leq L$.

The resulting spectrum usually contains much redundant information, such as fluctuations in frequency with limited spectrum of interest. Because of this need and the fact that some important distinguishing high frequencies are naturally attenuated by the vocal tract, a series of localized filters are applied to the spectrum in order to obtain an approximate equal resolution on the Mel-scale. The Mel-scale is an auditory scale that is similar to the frequency scale of the human ear. It is defined as

$$Mel(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.3}$$

Where '*f*' is the variable frequency. As the magnitude of each FFT, complex value is used to extract MFCC features, this process results in a scaled magnitude-frequency domain, which is down sampled by using a bank of triangular filters. For each filter, the magnitude coefficients are multiplied by corresponding filter gains and the results are accumulated as the amplitude

value. Mel frequency cepstral coefficients are then calculated from the log filterbank amplitudes by using the discrete cosine transform to reduce the spatial correlation between filter bank amplitudes

$$C_n = \sum_{k=1}^{K} S_k \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], n = 1,\ 2,\ 3.....L \tag{3.4}$$

where K is the number of log-spectral coefficients, S_k are the log-spectral coefficients and L is the number of cepstral coefficients required $(L \le K)$ [123]. For features extracted here, 12 coefficients plus a normalised log energy, or the *'zero'*th order cepstral coefficient are used. These coefficients form a ***13-dimensional feature vector*** for each frame.

When using hidden Markov models (HMMs) as the acoustic model, there is a fundamental assumption that the observations are conditionally independent. This requires removal of the temporal correlation of speech signals. Dynamic coefficients may be incorporated into the feature vector to reduce the temporal correlation [130]. These dynamic coefficients represent the correlation between static feature vectors of different time instances. One common form is the *delta coefficient* , Δo_t [123], which is calculated as a linear regression over a number of frames

$$\Delta o_t = \frac{\sum_{k=1}^{K} k\left(o_{t+k} - o_{t-k} \right)}{2\sum_{k=1}^{K} k^2} \tag{3.5}$$

Where k is the regression parameter, K is the width over which dynamic coefficients are calculated and the actual window size is 2k+1 accordingly. The second-order dynamic coefficients, *delta-delta coefficient*, $\Delta^2 o_t$, may also be calculated using a version of equation (3.5) in which the static parameters are replaced by the first-order delta coefficients. The ***26-dimensional acoustic feature vector***, i.e., length of the parameterised static vector (13 coefficients) plus the delta coefficients (+13), is used in this work.

In summary, the speech signal is divided into frame periods of 10msec.The FFT uses a Hamming window and the signal has the first order pre-emphasis applied using a coefficient of 0.97. The filter-bank has 26 channels and output 12 MFCC coefficients with a total of 26-dimensional acoustic feature vector.

3.3.2 HMM based Speaker Modelling

During training phase, speaker specific models are created using the feature vectors extracted by removing the irrelevant frames corresponding to non-speech or noisy data. During the verification phase, the same feature vectors extracted from the test utterance are matched against the relevant claimed model. Speaker modelling techniques are divided into generative and discriminative models. In generative models such as Gaussian mixture models or Hidden Markov models, the client speaker model is trained to maximize the likelihood of the data provided by the client without taking into account the impostor's information. In discriminative models such as artificial neural networks, the client model is trained to minimize the classification error between client and impostors, therefore, need some impostor data to create the client model.

Most text-dependent speaker verification systems use the concept of Hidden Markov Models (HMMs) that provide a statistical representation of the sounds produced by an individual [7]. The speaker models created by HMM depends heavily on the type of application, i.e., phoneme-level [120], word-level [124, 131], sentence-level, etc. The choice of HMMs in the context of text-dependent speaker verification is motivated by the inclusion of inherent time constraints and the topology of the HMM depends on the type of application. The standard ***left-to-right 5-state HMMs*** with five states per phoneme and three mixtures per state have been used for modelling in this dissertation.

Let O be a sequence of observed speech feature vectors corresponding to the HMM of a particular acoustic unit, for example, a word or a phone. It is defined as $O = [o_1, o_2, ...o_T]$ where o_t, $1 \le t \le T$, is a D dimensional feature vector and T is the length of the speech sequence. The generation process starts from the first non-emitting state. At each time instance, the state transits with a certain probability to either itself or the contiguous right state. The transition probability is a discrete distribution denoted as a_{ij} for transition from state *i* to state *j*. When an emitting state is entered, an observation is generated at that time instance with a probability density $b_j(o_t)$ for state j, which can be either discrete or continuous. Therefore, the observation sequence is associated with a state sequence, denoted as $S = [s_1, s_2, ...s_T]$.

In practice, only the observation sequence O can be observed and the underlying state sequence 's' is hidden. The observation sequence and the hidden state sequence are sometimes put together as $\{O, S\}$ and referred to as *complete data set*. The parameter set μ of an HMM, thus, includes the following parameters:

- a vector π of state distributions $\Pi = \{\pi_i \mid \pi_i = P(s_i = i)\}$
- state transition probability matrix, $A = \{a_{ij} \mid a_{ij} = P(s_{t+1} = j \mid s_t = i)\}$
- state output probability, $B = \{b_{jk} \mid b_{jk} = P(O_t = o_k \mid S_t = j)\}$, where $b_j(o_t)$ represent discrete probability distributions.

The Hidden Markov Model has three essential problems [132]:

3.3.2.1 Evaluation:

Given a HMM λ and an observation sequence $O = O_1, O_2, ..., O_T$, the probability $P(O \mid \lambda)$ of sequence *O* being generated by model λ is calculated using a forward-backward algorithm. The forward algorithm is based on the recursive computation of a forward probability $\alpha_t(i)$ and thus the probability $P(O \mid \lambda)$ is given as

$$P(O \mid \lambda) = \sum_{t=1}^{N} \alpha_T(i) \tag{3.6}$$

Where $\alpha_{t+1}(j) := \left[\sum_t \alpha_t(i) a_{ij}\right] b_j(o_{t+1})$ and $(1 \le j \le N; 1 \le t \le T-1)$

The backward probability $\beta_t(i)$ is the probability of having sequence *O* from time *t* + 1, with current state s_i for model λ and thus the probability $P(O \mid \lambda)$ is given as

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(o_i) \beta_1(i) \tag{3.7}$$

Where $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j),\ 1 \le i \le N$

3.3.2.2 Decoding

Given a model λ and a sequence of observations $O = O_1, O_2, ..., O_T$, the optimal path or most likely state sequence in the model that produced the observations is found using the Viterbi Algorithm. This algorithm is viewed as a special form of the forward-backward algorithm where only the maximum path at each time is taken into account instead of all paths. The most likely state sequence is obtained by backtracking of the optimal path for all times *t* (*t*= (*T*-1), (*T*-2)....1)

$$s_t^* = \psi_{t+1}\left(s_{t+1}^*\right) \tag{3.8}$$

Where $\psi_{t+1}(j) = \arg\max_i \left[\delta_t(i) a_{ij}\right]$ and $\delta_{t+1}(j) = \max_i \left[\delta_t(i) a_{ij}\right] b_j(o_{t+1})$;

3.3.2.3 Estimation

Given a model λ and a sequence of observations $O = O_1, O_2, ..., O_T$, the model $\{\pi, A, B\}$ that maximizes $P(O \mid \lambda)$ is trained using expectation maximization (EM) algorithm. The Baum-Welch algorithm is considered a special case of the EM algorithm and is used for training the HMM models. As the optimization criterion, this algorithm uses the total production probability $P(O \mid \lambda)$. The probability of model being in state s_i at time t, given O and λ in terms of α and β is given as

$$\gamma_t(i) = P(S_t = i \mid O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{P(O/\lambda)} \tag{3.9}$$

The updated models are re-estimated using the below expressions:

$$\hat{\pi}_j = \gamma_1(i) \tag{3.10}$$

$$\hat{a}_{ij} \frac{\sum_{T=1}^{T-1} \gamma_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.11}$$

$$\hat{b}_j(o_k) = \frac{\sum_{t:O_t=o_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{3.12}$$

After the re-estimation, if new model parameters are more likely than the old ones, i.e., $P(O \mid \hat{\lambda}) \geq P(O \mid \lambda)$, then the new model will replace the old one. The re-estimation is continued until the above condition is valid.

For speaker verification tasks with a small vocabulary, such as digits, HMMs are often used to model individual words. However, for verification with medium to large vocabularies, it is difficult to obtain sufficient training data for each individual word in the vocabulary. The most commonly used solution is to use HMM to model sub-word units such as phones. Each word can be easily split into a sequence of phones with each phone considered as a sub-word unit. There are two main types of phone model sets i.e., context-independent phones (mono-phones) and context-dependent phones [123]. The *mono-phone* set does not take into account the context information, i.e., the dependence of phone pronunciation on the preceding and following phones. To model these variations, *context-dependent phones* such as **tri-phones** are used. For example, consider the phone *ah*, a possible triphone may be 'w-ah+n', where 'w' is the preceding phone and 'n' is the following phone, '-' denotes the preceding (left) context and '+' denotes the following (right) context. Therefore, the tri-phones for an isolated word "*one*" are [w+ah w-ah+n ah-n].

The issue with using tri-phones is that the number of possible acoustic units is significantly high and so difficult to collect sufficient data to train all tri-phones robustly. One solution for this issue is the parameter tying, or clustering technique [133] that considers a group of parameters as sharing the same set of values. Tying can be performed at various levels, such as phones, states, Gaussian components, or even mean vectors or covariance matrices of Gaussian components. The most widely used approach is to do state level parameter tying, referred to as ***state clustering***.

The phonetic decision tree approach is used to efficiently perform state clustering [133] for rare and unseen context. Clustering is performed in a top-down fashion with a set of questions related to the left and right contexts of each phone. At the beginning, all states are grouped to root node and later split into children nodes based on the answers for these context questions. This split process stop when the amount of training data associated with the current node falls below a minimum threshold. Though decision tree clustering is a local optimal binary search, it can efficiently handle the problem of unseen tri-phones as all contexts are mapped to a leaf node. Hence, it is the most popular state clustering approach and is adopted in this work.

3.3.3 Speaker Model Adaptation

Though the trained models accommodate a wide range of acoustic variability, there always exists a mismatch between the conditions in which models are trained and those in which they are used. Different microphones, transmission channels, background noises or speaker characteristics can introduce this mismatch. However, the models could be adapted to new conditions by modifying the parameters using a small amount of data. One typical situation is the case of speaker variability. Several studies on unknown [81, 114] and text-dependent [134] speaker recognition tasks have demonstrated the effectiveness of this technique. A speaker-dependent system can offer significant WER reductions in comparison to a speaker-independent system when sufficient data is available to retrain the system [135], but speaker adaptation can give almost the same results with a reduced amount of speaker-specific data.

Speaker model adaptation can be either supervised or unsupervised. Supervised adaptation is the process in which the transcriptions of the data are provided. Adaptation is performed usually by asking the user to pronounce a given set of sentences that are used in turn to adapt the acoustic models. Second method is the unsupervised adaptation that can track non-stationary mismatches. However, the unsupervised adaptation requires a good match between target speaker model and testing utterance to adapt the speaker model. However, the performance results may be imperfect yielding incorrect transcriptions. The adaptation techniques can be divided into three modes based on the amount of available adaptation data [136].

- Batch Mode: The adaptation is performed off-line and sufficient adaptation data are necessary.
- Incremental Mode: The adaptation is done successively based on the speech from the speaker. This mode updates the model parameters and discards the used data periodically, which can reduce the computation and memory requirements efficiently compared with the batch mode.
- Instantaneous Mode: The adaptation uses the same utterance to be verified and so the process of adaptation and verification are performed at the same time.

The model adaptation techniques are also classified into transformation-based (indirect) and Bayesian (direct) adaptation. The transformation-based approach separately transforms clusters of HMMs according to their transformation functions. The proposed

techniques for transformation-based adaptation include maximum likelihood linear regression (MLLR), constrained transformation, and maximum likelihood stochastic matching (SM). The Bayesian adaptation approaches include maximum a posteriori (MAP) and quasi-Bayses algorithm. The ***Maximum Likelihood Linear Regression (MLLR)*** [137] and ***Maximum A Posteriori (MAP)*** [138] methods are widely used in literature and also in this dissertation work. These methods are described below:

3.3.3.1 Maximum Likelihood Linear Regression (MLLR)

The most successful method for the fast adaptation of HMM parameters on very limited data is the maximum likelihood linear regression (MLLR) [137]. This method computes a set of transformations that will reduce the mismatch between an initial model set and adaptation data. The transformation matrix used for a new estimate of the adapted mean is given by $\hat{\mu} = W\xi$ where *W* is the $n \times (n + 1)$ transformation matrix and ξ is the extended mean vector given as:

$$\xi = \left[w\mu_1\mu_2 ... \mu_n \right]^T \tag{3.13}$$

Here, *w* represents a bias offset whose value is fixed at one. Hence W, *transformation* matrix can be decomposed as $W = \left[b\ A \right]$ where A represent $n \times n$ transformation matrix and b represents a bias vector.

3.3.3.3 Maximum A Posteriori (MAP)

Speaker model adaptation with the maximum a posteriori (MAP) approach is based on the use of prior knowledge about model and adaptation data to obtain an adapted model. The speaker model obtained in training phase is used as a base for the new model and the parameters of prior distributions are updated to obtain an adapted general model that covers the newly observed data. This adaptation process is sometimes referred to as Bayesian adaptation.

In MAP adaptation, a posterior probability that the model λ matches the observation *O* is maximized by updating the parameters of an initial model [139]**.** The adapted model is defined as:

$$\lambda_{MAP} = \arg\max_{\lambda} P(\lambda \mid O) \tag{3.14}$$

The new adapted formula using Bayes' rule is given as:

$$\lambda_{MAP} = \arg\max_{\lambda} \frac{P(O|\lambda)P(\lambda)}{P(O)} \tag{3.15}$$

Where $P(O|\lambda)$ is the likelihood probability of the observation sequence *O* given the speaker model λ, $P(O)$ is the prior probability of observing O, $P(\lambda)$ is the probability density obtained from the pdf of an initial model. The update formula for adaptation of state *j* and mixture component *m* is defined as:

$$\tilde{\mu} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm} \tag{3.16}$$

$$N_{jm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L^r_{jm}(t) \tag{3.17}$$

Where τ is a weighting of the a priori knowledge to the adaptation data, μ_{jm} is the speaker independent mean, $\bar{\mu}_{jm}$ is the mean of the observed adaptation data, $\hat{\mu}_{jm}$ is the mean of the adapted model and *N* is the occupation likelihood of the adaptation data.

$$\bar{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L^r_{jm}(t)o^r_t}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L^r_{jm}(t)} \tag{3.18}$$

As a result, when the likelihood of occupation of a Gaussian component (N_{jm}) is small, the estimated mean MAP is close to the model mean. With MAP adaptation, every single mean component in the system is updated with a MAP estimate, based on the *a priori* mean, the weighting and the adaptation data. Hence, MAP adaptation requires a new *'speaker-dependent'* model set. A drawback of this approach is that more data is necessary for an effective adaptation when compared to MLLR.

The two adaptation processes can be combined to improve performance, by using the MLLR transformed means as the priors for MAP adaptation. The advantages of two adaptation techniques can be combined to form the hybrid adaptation approach [140]. With the increased amount of adaptation data, there is not only a set of global MLLR

transformation functions for rapid adaptation, but also locally modified model parameters by MAP when large amounts of adaptation data are available.

3.3.4 Decision Making

The speaker verification task involves making a decision whether the test data provided during the verification phase belongs to the claimed model or not. After computing a match score between the test data's feature vector and a model of the claimed speaker's voice, a verification decision is made to accept or reject the speaker's claim. Accept or reject decision process is nothing but to accept or reject the hypothesis of the testing problem. Given a speech segment *X* and a claimed identity *S* the speaker verification system should choose one of the following hypotheses:

H_S: *X* is pronounced by *S*

$H_{\bar{S}}$: *X* is not pronounced by *S*

The decision between the two hypotheses is usually based on a likelihood ratio given by

$$L(X) = \frac{p(X \mid H_S)}{p(X \mid H_{\bar{S}})} \begin{cases} > T \ accept \ H_S \\ < T \ accept \ H_{\bar{S}} \end{cases} \tag{3.19}$$

Where $p(X \mid H_S)$ and $p(X \mid H_{\bar{S}})$ are the probability density functions or the likelihoods associated with the speaker *S* and non-speaker $\overline{S}$, respectively. '*T*' is the threshold to accept or reject H_S. In practice, H_S is represented by a hypothesized speaker's model λ_S, and $H_{\bar{S}}$ is represented by a model $(\lambda_{\bar{S}})$ estimated using a set of other speakers that cover as much as possible space of the alternative hypothesis.

There are two main approaches to selecting this set of other speakers. For each claimed speaker S a set of speakers $\overline{S_1}, \overline{S_2}, \overline{S_3}, ..., \overline{S_N}$ called a cohort set of speakers [124] can be selected that are representative of the population near the claimed speaker. In this case, each speaker will have a corresponding non-speaker model. Another way of choosing the cohort speaker set is to use speakers that are typical of the general population. Reynolds [141] reported a more practical approach that is modelled using randomly selected, gender-balanced background speaker population outperformed a population near the claimed speaker. This model is usually trained using speech samples from a large number of

representative speakers. This model is referred in the literature as World Model or Universal Background Model (UBM) [142]. This last approach is the most commonly used in speaker verification systems. It has the advantage of using a single non-speaker model for all the hypothesized speakers, or two gender-dependent world (background) models. The likelihood ratio in (3.19) is then rewritten as

$$L(X) = \frac{p(X \mid \lambda_S)}{p(X \mid \lambda_{\bar{S}})} \tag{3.20}$$

Often the logarithm of this ratio is used. The final score is then:

$$S(X) = \log L(X) = \log p(X \mid \lambda_S) - \log p(X \mid \lambda_{\bar{S}}) \tag{3.21}$$

Once the model is trained, the speaker verification system should make a decision to accept or reject the claimed identity by comparing the score of the test utterance with a decision threshold. The speaker verification performance is affected by the selection criterion and the value of a threshold used for making the decision. The threshold is usually chosen during the tuning or development phase, and could be either speaker-dependent or speaker-independent. The threshold estimation, in general, does not account for the intra-speaker variability and the mismatch between tune and test data conditions. Therefore, the optimal operating point could be different from the pre-set threshold. There are two main approaches to dealing with the problem of threshold estimation. The first one consists of setting a priori speaker-dependent threshold [143], in such a way that the threshold is adjusted for each speaker in order to compensate the score variability effects. The other approach is to use score normalization techniques [144] to make the threshold more robust and easy to set. Some of these issues specific to text-dependent speaker verification such as threshold selection and background model design are discussed in the next section.

3.4 Issues with Text-dependent Speaker Verification

The design of a successful text-dependent speaker verification system is a highly complex procedure and involves many challenges and issues [145]. These issues can be related to either the technological or the deployment aspect of the system design. Most of these issues are the sources of error and thus reduce the overall system performance.

The issues with technological aspects of the system are mainly concerned with the core algorithms (e.g., feature extraction, matching) and/or fusion techniques that can be employed for performance enhancement. For example, recent work has shown that great improvement can be achieved in speaker verification accuracy through information fusion approaches. Campbell et al. [146] have proposed the linear combination of classifiers based on a variety of feature types, including acoustic, phonetic, prosodic and even lexical ones for improvement in performance. Sanderson et al. [147] have proposed a hybrid fusion approach in which MFCC, CMS and MACV features are fused using adaptive decision fusion. Kevin [148] presented a text-dependent speaker verification system that uses data fusion concepts to combine the results of classifiers based on dynamic time warping (DTW) and the neural tree network (NTN). The use of linear opinion pools has shown an equal error rate of 2% which is better than the individual performance of either classifier.

The deployment of a text-dependent speaker verification system is a difficult task and the related issues involve the determination of system parameters such as amount of speech required for accurate training and testing. Another challenge is the estimation of a threshold for acceptance and rejection and the eventual error rate to expect from such a threshold. This selection is usually difficult in the presence of a significant mismatch between the tune/development dataset and the test dataset conditions. The deployment issues that are to be addressed for the proposed system architecture are:

3.4.1 Limited Data and Constrained dictionary

The performance of a verification system depends mainly on the amount of speech data available for training a model and the length of test speech data. The evaluations on NIST-SRE [149] have confirmed that the duration and number of sessions of training and verification affect the performance of speaker verification systems. The text-dependent applications, in general, use multiple repetitions of data from multiple sessions for training (e.g., 4-8s utterances) whereas single utterance from a subset of the training text is used for testing (2-3s utterances). These requirements are shown in [7] to be motivated by the usability studies which show that shorter sessions for training and testing provide greater user convenience and so generally preferred by end customers.

To achieve reasonable accuracies using short training and testing constraints, the dictionary needs to be strictly restricted. The general examples of the dictionary used for training include the individual digits from *zero* to *nine*, digit telephone numbers, account

numbers, first and last names, or speaker's voice as a password. In most cases, the text chosen for testing from dictionary exactly match the text used during training. The results presented in [7] also shows that the performance is better for the conditions where the training and testing dictionary is similar (for example, 5.05% for 9-digit account numbers and 6.16% for individual digits). The testing dictionary, however, could be a random combination of digits selected from training dictionary. This method allows for a longer verification utterance without increasing the relative load. For example, in the case of random digit sequences, the digits in the training and testing dictionary may not be same and so the performance (11.5%) is lower compared to scenarios where the same text (i.e., individual digits and 9-digit number) from dictionary is used for training and testing.

The mismatch in data used for training and testing can result in performance degradation, for example, in [150], the equal error rate is shown to increase by a factor of 5 for mismatch conditions. It is, therefore, significant to minimize these mismatches in training and testing data/text to avoid a significant degradation in verification performance. These issues are considered in the design of the experimental protocol used for experimental evaluation of text-dependent speaker verification.

3.4.2 Length of the speech data

The length of enrolment and testing utterances has a significant effect on the verification performance. Jason et al. [151] investigates the effect of varying enrollment and test utterance lengths on score distributions and consequently their effect on performance. In addition, the study examines the mixing of a number of variable training length utterance trials in a single evaluation. Similar studies that demonstrate the effect of utterance length on text-dependent speaker verification performance are listed in table 3.1. Although the features and methods used for verification are different, the verification performance often improves with an increase in the length of the train and/or test utterances.

Most of the speaker verification systems use an entire utterance to arrive at a decision to accept or reject the claimed identity of the speaker. In many practical applications, especially in defence and intelligence, it is desirable to make a decision while the speaker is talking. In such scenarios, data is usually made available in a stream or small chunks. It would be beneficial to make sequential decisions on the smaller blocks of data as they become available rather than wait for the entire utterance. The most general sequential decision strategies that can be used are heuristic method [152] or methods based on Wald's

sequential probability ratio (SPRT) test principle [153]. The SPRT approach enables to demonstrate that accurate verification decisions are obtained after only 2-10 seconds of evaluation data where usually 100 seconds are needed [93]. Similarly, in [21], it is shown that final decision from combining sequential decisions of seven digits per utterance is as reliable as using a fixed length of 10 digits, thereby reducing the computational cost by 30%. This method is thus shown to systematically trade-off between the performance of the system and the amount of data needed to make a decision. However, the analysis does not consider the order in which the decisions are combined which is significant in sequential fusion (as the fusion decision at a stage is dependent on the preceding stage decision). Kato and Shimizu [131] studied the significance of preserving the sequence of digits for performance improvement of a text-dependent system. It is demonstrated that a relative improvement of more than 50% is achieved when the digit sequence used in the testing preserves the order of data sequence collected during training.

In the next chapter, the sequential decision fusion approach is evaluated for the trade-off between performance and number of decisions required to make a reliable final decision when the sequence of the instances (either digits/words) is the same for training and testing.

3.4.3 Intra-speaker Variability

The efficiency of speaker verification systems mainly depends on inter-speaker variability, i.e., the variation of speech between different speakers. The greater the inter-speaker variability between true speaker and impostor, the more accurate a system is likely to be. The system performance also depends on intra-speaker variability that refers to the variation in speech from a single speaker, i.e., a speaker cannot reproduce the same word or phrase in exactly the same way. Different speaking rates, emotional state and speaking environment (e.g. speaking against background) cause the intra-speaker variability noise. The latter, known as The Lombard effect, is the tendency to increase one's vocal intensity and to modify intonation when speaking in a noisy environment.

The speaker models trained using limited speech data may not be representative of the all speaker characteristics because of these intra-speaker variability's, thereby affecting the speaker verification performance. In addition to the above-mentioned factors of variability, the physiological changes such as natural aging and behavioral changes can also affect

Table 3.1 Text-dependent speaker verification results for different feature extraction and modelling techniques

Reference	**Features**	**Method**	**Error Rates in % (length of utterance in seconds)**
Linares et al. 1998 [154]	Mel-cepstrum coefficients	GMM	8 Digits FAR - 3.7, FRR - 8.1
Wong et al. 2001 [155]	MFCC	HMM	6 Digits - 0.57
Javier et al. 2004 [156]	MFCC	GMM	8 Digits - 9.6
Allano et al. 2006 [157]	MFCC	GMM	5 Digits – 7.21 10 Digits - 3.24
Subramanya et al. 2007 [158]	MFCC	HMM - Likelihood Ratio Test	2 Digits – 3.35 4 Digits – 1.89 6 Digits – 0.63

the system performance. As an example, first-time users of a speech application (usually the training session) tend to cooperate with the system by speaking slowly under friendly conditions. As these speakers get more exposure to the application, they can alter the way that they interact with it and use it in adverse conditions (such as different channels) that results in a decrease in accuracy. To deal with this problem, incremental enrolment techniques are used to include the short and long-term evolution of the voice [114].

3.4.4 Background Model Design

Virtually, all state-of-the-art speaker verification systems use background models to enhance the robustness and computational efficiency of the verification system. In the training phase, the target models are adapted from the background model [141], whereas in the verification phase, background speakers are used in the normalization of the speaker match score [116, 141]. Therefore, the design of background models is crucial to the accuracy of a speaker verification system.

The background models can be trained on a set of speakers called as either likelihood ratio sets [116], cohorts [124] or background sets [159]. The selection, size, and combination of the background speakers have been studies in detail [124, 159, 160]. In general, to obtain the best performance the background models require the use of speaker-specific background speaker sets. In [159], an algorithm for the selection of background speakers for a target user is developed for a text-dependent task. However, the use of speaker specific sets can be a drawback in applications using a large number of hypothesized speakers, each requiring their own background speaker set.

Alternatively, a universal background model [142] can be trained on a pool of speech samples from several speakers representative of the population of speakers expected during verification. The main advantage of this approach is that for each particular task, a single speaker-independent model can be trained and then used for the speaker verification. The verification based on a Universal Background Model can be improved by selecting the UBM according to the dependencies of the task/application such as the noise, gender or recording conditions. For example, it is shown in [125], that using handset-matched background models reduces false acceptances (at a 10% false rejection rate) by more than 60% over the handset-independent approaches. Similarly, the dependency on text/lexical content for background model selection have a significant positive impact on performance [161, 162]. Therefore, the background models in this dissertation are mostly trained for each specific text/word (text-dependent).

3.4.5 Channel Variability

The performance of the verification system depends on the type of the channel used for collecting the speech samples. The two primary channels encountered for speaker verification are microphones and telephones. The quality of the microphone used in speaker verification systems will also have an effect on performance. Some microphones are more susceptible to noise and each has a different *Signal to Noise* Ratio. An omni-directional microphone, for example, has a uniform pickup pattern and consequently, is likely to be most susceptible to ambient noise. In contrast, directional microphones are designed to respond to sound from a single direction (unidirectional microphones) or from two specific directions (bidirectional microphones). Because they can isolate sound sources based upon location, directional microphones are better for speaker verification than omni-directional

microphones. Das et al. [163] showed that the use of different microphones for training and testing significantly increases the error rate.

Speaker verification applications over a telephone must accommodate the noise generated by microphones in the telephone handset. If the system is on the telephone network rather than using a handset, the system is susceptible to noise in the telephone network. There are different types of microphones used as standard telephone equipment, such as Electret handsets, Carbon button handsets. These classes of microphones vary in amount of reverberation captured from the environment and their patterns of distortion. In [129], the additional information about the telephone channel is explained.

Irrespective of channel type, the conversion of input signal results in patterns of signal distortion introduced by the channel. The distortion is a by-product of the acoustic patterns and capabilities of the input device. The channel also contributes its own additive noise to the signal, usually electrical noise. In cases where the training speech data is collected over the same type of channel and handset microphone, the distortion effect is roughly the same. Consequently, unless the distortion is particularly acute, the comparison between speaker model and speech signal is not usually adversely affected. If, on the other hand, the training speech data is collected on a different type of channel from the subsequent verification speech, then performance does degrade. There are several approaches to address this problem. The best and most effective approach is for the system to employ multiple models for each speaker (i.e., one model for speech from different channels), which in general, is not the most practical solution. Alternatively, speech from the various channel sources can be incorporated into a single model. The stochastic modelling approach facilitates the incorporation of mixed channel data, to a certain extent, using channel compensation techniques.

The channel mismatch in recording conditions between training and testing is the main challenge for speaker verification. Differences in the background noise, in the telephone handset or microphone, in the transmission channel and in the recording devices can introduce variability over the recording and decrease the accuracy of the system. This is mainly due to the statistical models that capture not only the speaker characteristics but also the environmental characteristics. The system decision can, therefore, be biased if the verification environment is different from the training. The features and score normalization techniques [112, 141] are useful to make speaker modelling more robust to recording conditions. The use of high-level features [164] that are more robust to mismatched

conditions can be used for performance enhancement. The verification in this dissertation considers speech samples from the same channel type for the training and testing conditions. Score normalization techniques are used to compensate for the channel mismatches occurred during data acquisition.

3.4.6 Threshold Estimation Criteria

The determination of decision thresholds is another important issue for text-dependent speaker verification. Conventional threshold determination methods [83, 165] typically compute the distribution of inter- and intra-speaker distortions, and then choose a threshold to obtain an equal error rate (EER) or a pre-defined error rate, i.e., false acceptance rate (FAR) and false rejection rate (FRR). The success of this approach, however, relies on whether the estimated distributions match the speaker- and impostor-class distributions.

The use of an impostor model for speaker score estimation enables accurate and easy decision threshold estimation and improved speaker separability. The methods such as likelihood normalization [166], cohort normalized scoring [124], and minimum verification error training [167] help to select an appropriate threshold but may cause the system to favour rejecting true speakers, resulting in a high FRR [116]. The estimated thresholds can be either speaker-independent or speaker dependent. A speaker-dependent threshold better reflects speaker peculiarities and intra-speaker variability than a speaker-independent threshold.

The performance of the system is shown to be enhanced by adapting speaker-dependent thresholds [109]. Session-to-session speaker variability, however, contributes much bias to the threshold, rendering the verification system unusable. Due to the difficulty in determining a reliable threshold, researchers often report the equal error rate (ERR) of verification systems based on the assumption that a posteriori threshold can be optimally adjusted during verification. However, a report [168] based on a threshold estimation using normalization techniques has found that the average of FAR and FRR is about 3 to 5 times larger than the ERR, suggesting that the ERR could be an over optimistic estimate of the true system performance. As EER is less significant, in real cases, a certain FRR or FAR is usually required for threshold setting.

The criteria for threshold selection of base classifiers can be based on Equal Error Rate for each classifier [169]. The other criteria could be to Weighted Error Rate (WER) that enables to obtain either Equal FRR or Equal FAR on multiple classifier systems. As the

variation in threshold setting criteria influences the performance, multiple threshold selection criteria are evaluated and then the most appropriate criterion with minimum verification errors is selected. In addition, speaker-dependent and digit dependent thresholds are used for improvement in text-dependent speaker verification.

3.4.7 Protection against Spoof Attacks

Speaker verification systems are susceptible to some level of spoofing that either involves the playback of static phrase or an impostor trying to impersonate the speaker. An impostor can replay the client speaker's recording in order to be falsely accepted by the verification system as the legitimate client. This relay attack, however, requires a high-quality recording of the client's utterance and digital equipment to perform the playback. Further, the impostor will have to be able to record the client's voice beforehand in order to gain access to the system. The recording could be done in general or during a transmission attack that begins with the passive eavesdropping attack on the communication channel. The impostor can later 'replay' the captured data with legitimate credentials to the verification system resulting in a false acceptance of the attacker/impostor. The protection against recordings is important, especially, for text-dependent speaker verification systems. If the system is purely text dependent and an impostor has gained access to a recording, it becomes relatively easy to fool the verification system [170]. However, the advantage of using speech for verification is that it is natural to prompt for different combination of the enrolment sequences.

An impostor with access to recordings of a client speaker can use different transformation techniques to alter his/her voice to sound like the target client speaker. The speaker verification systems are also vulnerable to these altered imposter voices [171] and so impostor acceptance can be increased [172]. However, these techniques usually require technical expertise and complete knowledge of the target speaker verification system (such as feature extraction, modelling method, background model, target speaker model, and other algorithms). In most cases, it is difficult for an impostor to obtain this information because implementation details and internal algorithms of security systems are usually kept secret for commercial speaker verification systems.

The replay attack can be counter measured using the liveliness assurance, i.e., prompting the speaker for a random utterance. A replay attack using data from transmission recording can be prevented by incorporating current time into the information submitted to verification system (this method is known as "time-stamping"). For time stamping to be

effective the time must be incorporated in such a way that an impostor cannot capture the data as before, modify the time field, and replay the data. The other method proposed for protection against recordings is to perform continuous speaker verification of the system. The attacks by natural mimicking or voice conversion can also be counter measured by prompting the speaker for multiple texts (words or phrases) thereby reducing the chances for an impostor acceptance (as it might be difficult for an impostor to reproduce the target client's voice for multiple words/phrases).

Most of these issues are addressed while designing the text-dependent speaker verification. For example, the databases used for empirical evaluation have multiple utterances for the same text thereby ensuring better modelling of intra-speaker variability. The mismatch between the training and testing sessions is minimised in most cases by considering the training data from various sessions and score normalisation techniques. Speaker dependent and text-dependent thresholds are used for performance improvement of the verification system. The three threshold selection criteria, i.e., Equal Error Rate for each classifier, Equal FAR or Equal FAR for different classifiers, are evaluated to find the criterion with best performance of the proposed fusion scheme. Same dictionary is used for training and testing to reduce the degradation in performance. These addressed issues are discussed, in next section, while explaining the databases and protocols used for empirical evaluation of text-dependent speaker verification.

3.5 Experiment Design

3.5.1 Database

One key element for evaluation of speaker verification performance is the availability of speech databases. Most databases require speech samples from a large population of individuals, together with the desirable presence of possible factors of speech variability (i.e., multi-session, multiple languages, multiple environmental conditions, etc.). The overview of current publicly available corpora for speaker recognition evaluation are presented in [173]. Some of the English databases used for text-dependent speaker verification are YOHO, POLYCOST and TI-46 Word. These databases have multiple repetitions of utterances collected in different sessions. However, in most of these databases, the words/phrases collected are mostly speaker dependent (e.g., 7 digit client codes specific to a user in POLYCOST) or the number of utterance repetitions are not sufficient to perform verification

at reasonable significance (e.g., TI-46 Word database has just 26 utterances which are insufficient for training and performing multi-sample fusion testing). The experiments in this dissertation are evaluated using speech data from CSLU [174] and AVICAR [175] databases.

3.5.1.1 CSLU Speaker Recognition Database

The Centre for Spoken Language Understanding (CSLU) Speaker Recognition corpus (formerly known as Speaker Verification) [174]**,** consists of telephone speech from 91 participants. The speech data for each participant are collected in twelve separate sessions over a two-year period. The data in this corpus are collected over digital telephone lines and recorded with the CSLU T1 digital data collection system. The protocol of this database is designed for the use of vocabulary dependent and vocabulary independent speaker verification systems. The corpus is also designed to provide sufficient data to study variability within and across sessions. Hence, several different types of data are requested from each speaker during data collection [176]. The categories of data that are suitable for the proposed text-dependent speaker verification architecture are digit stings, words and phrases (table 3.2). These categories have four repetitions of each text repeated over twelve sessions for each speaker (i.e., nearly 48 utterances).

Table 3.2 A few categories of data recorded from the CSLU speakers

Category	**Description**	**Examples**
Single Words	Words selected for their phonetic coverage	*mango, button, choices, decision, whereabouts, azure, offstage, little*
Phonetically Rich Phrases	Phrases generated for phonetic combinations that are not frequent in American English.	*"stop each car if it's little",* *"play in the street up ahead",* *"a fifth wheel caught speeding", etc.*
Digit Strings	Short and randomly generated digit string	*AP -5 3 8 2 4, AQ - 6 1 oh 9 7* *AR - 4 zero 7 1 3, AS - 2 8 3 7 6* *AT - 1 9 0 5 4, AU - zero 5 2 3 9*

3.5.1.2 AVICAR (Audio-Visual speech In a CAR)

The other database used for evaluation of the proposed fusion scheme is AVICAR database [175] that is an in-car speech corpus containing multi-channel audio and video recordings. The speech recognition train and test database is recorded in a moving automobile using an array of four cameras and eight microphones. The database includes audio and video data from nearly 87 subjects. Each recording session contains speech under five noise conditions from common driving scenarios, i.e., IDL (Engine running, car stopped, windows up), 35U and 35D (Car travelling at 35mph, windows up and down respectively), 55D and 55U (travelling at 55mph, windows up and down respectively). This enables to analyse the effect of different types of noise on verification performance.

Ten different script sets are used for the corpus with only isolated digits (1, 2, 3... 9, 0, oh, zero and done) and letters (a, b, c... z) common for all the speakers. Subjects are asked to speak isolated digits and letters twice under each noise condition. Therefore, 70 utterances (from all the five noise conditions) are collected for each digit and letter for a speaker. Text-dependent speaker verification is performed by modelling each digit and letter as a separate HMM model.

3.5.2 Experimental Protocol

The most commonly used vocabulary for speaker verification is digits, words and phrases. These three categories usually present similar (if not identical) security concerns. However, the clear preference from speakers, in general, is digits as people are used to remembering phone numbers and short digit strings (four to six digits) that can be randomly generated. The typical implementation of a text-dependent system is the use of speaker digit information (such as the date of birth, a telephone number, PIN, account numbers or SSN) to determine the identity claim of the individual. During enrolment, digit dependent speaker-specific models can be trained for different digit utterances of a speaker. The advantage of using digits for verification is that only few (generally 10) models are to be trained. During verification, a random digit sequence is selected and the speaker is verified for the corresponding classifier models. By combining digits randomly, the liveliness of the speaker can also be ensured. Further, it is possible to use a secret numeric PIN specific to a speaker to enable additional security. Therefore, the speaker verification evaluation in this dissertation is performed on isolated digits. The digits 1, 2, 3, 4, 5, 7 and 9 have three repetitions whereas

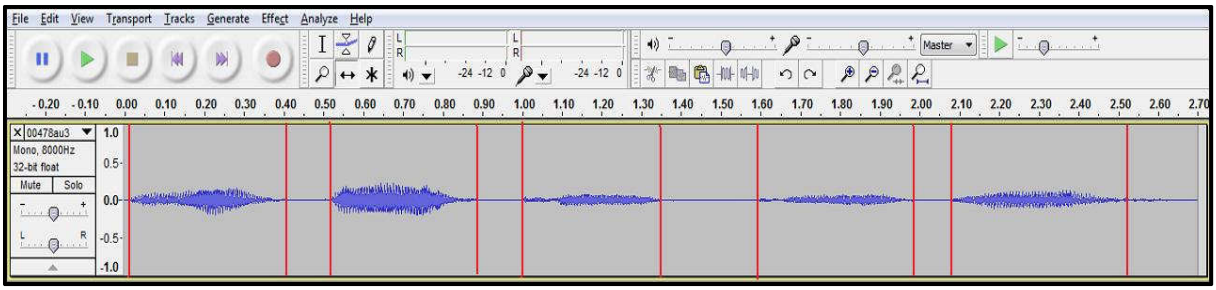

Figure 3.2 The Audacity Software [177] screen used for manual segmentation of digit strings from CSLU database (Digit String - *'Zero-five-two-three-nine'*)

digits 6 and 8 have two repetitions in the six digit-strings.

The AVICAR database has isolated digit data but from different noise conditions. The CSLU corpus, however, has clean speech data but in the form of digit-strings (i.e., a sequence of 5-digits). The isolated digits for CSLU database are obtained by segmenting the digit-strings into individual digits. This segmentation is performed manually using Audacity software [177]. Figure 3.2 shows example of digit string "*Zero-five-two-three-nine*" and its manual segmentation using Audacity. The number of utterances used for performance evaluation is increased by segmenting the digit-strings in CSLU database. For example, the digit *'one'* has three repetitions in the digit strings (AQ, AR and AT) and so speaker verification can be performed on 144 utterances that are subdivided for training and testing.

The evaluation of the proposed architecture depends on both multi-instance and multi-sample fusion schemes. Multi-instance fusion is based on the combination of decisions from different digits (here each ***digit*** is considered as an ***instance***) and so it is significant to analyse the results for multi-instance fusion at reasonable precision. As the digits zero, oh, six and eight have insufficient data (less than 100), the remaining digits are used for initial performance evaluation. In addition to the speech data, the protocol used for dataset division is also significant for performance evaluation.

The evaluation is performed by dividing the speech data into three different data sets, i.e. train, tune/development/validation and test datasets. The ***train dataset*** has utterances that are used in modelling text-dependent HMMs that are specific to each individual speaker's characteristics. The ***tune dataset*** has utterances that are verified using the trained HMM models for each speaker. The results from these verification tests are used to fine-tune the parameters required for verification, for example, to estimate the decision thresholds,

determine the number of classifiers used for fusion, etc. The parameters selected on tune dataset are used for text-dependent speaker verification on the ***test dataset.*** It is significant to ensure that these datasets are disjoint to each other so that the resultant performance of the system is not overestimated.

3.5.2.1 SET-*1*

For speaker verification testing, the data selected are obtained from only the male speakers and those with a complete set of utterances for six digit-strings. As the manual segmentation of the digit-strings is time consuming, only a selected number of (random) speakers data are used for testing. Although the size of the selected data is small (segmented digits for 11 speakers), the utterances used for testing are of clean speech. The main objective of the thesis is to determine if the proposed architecture enables a better control over the trade-off between verification error rates. Therefore, the pattern of increase/decrease in each type of error for different fusion schemes is of greater significance than the actual variation in errors. The objective can thus be validated irrespective of the data size used for testing. However, the results obtained for this dataset are used for initial evaluation and the size of the dataset is increased (using conversion tools) later to perform evaluation with greater precision.

The protocol explained above is used for segmenting the speech data for training, tuning and testing a speaker from the CSLU Database. These tests are performed on 11 random male speakers and the data is referred to as ***SET-1***. Each speaker's data is divided into three disjoint datasets (train, tune and test sets) for each digit (1, 2, 3, 4, 5, 7 and 9). Impostor testing for a speaker (client) is performed using data from the 10 speakers other than the client. The verification performance for each speaker is analyzed on four different training models (one utterance from a session for each model) to ensure accurate results with reliable precision. Each train dataset has utterances from different sessions and digit-strings to better model the intra-speaker variability. For example, a train dataset for 'digit 1' is a combination of one segmented utterance from each of the digit-strings AQ, AR and AT for seven different sessions (1 utterance * 3 digit-strings * 7 sessions). The remaining data (other than train dataset) is divided into four different combinations of two disjoint datasets, i.e., the development and test dataset. The selection of the tune dataset for each train dataset is based on the digit-string from which a particular digit is segmented. For example, the tune datasets for 'digit 1' are selected based on:

Tune Dataset 1: subset of segmented data from the digit string AQ (6 1 *oh* 9 7)

Tune Dataset 2: subset of segmented data from the digit string AR (4 *zero* 7 1 3)

Tune Dataset 3: subset of segmented data from the digit string AT (1 9 *oh* 5 4)

Tune Dataset 4: subset of segmented data from the combination of strings AQ, AR and AT

The remaining data (other than train and tune datasets) is used for testing the performance of text-dependent speaker verification. For each digit, four speaker specific digit dependent HMM based models are trained on 21 client utterances. For each train dataset, the remaining data is divided into four combinations of development and test datasets. Each development dataset has 35 client and 140 impostor utterances for fine-tuning the parameters, such as threshold estimation or determination of base classifier performance. The test dataset, which is disjoint to the corresponding train and development datasets, has 70 client and 420 impostor utterances for determining the base classifier performance.

The performance of base classifiers is usually determined by selecting an optimal threshold on development/tune dataset for each digit. The two commonly used criteria for

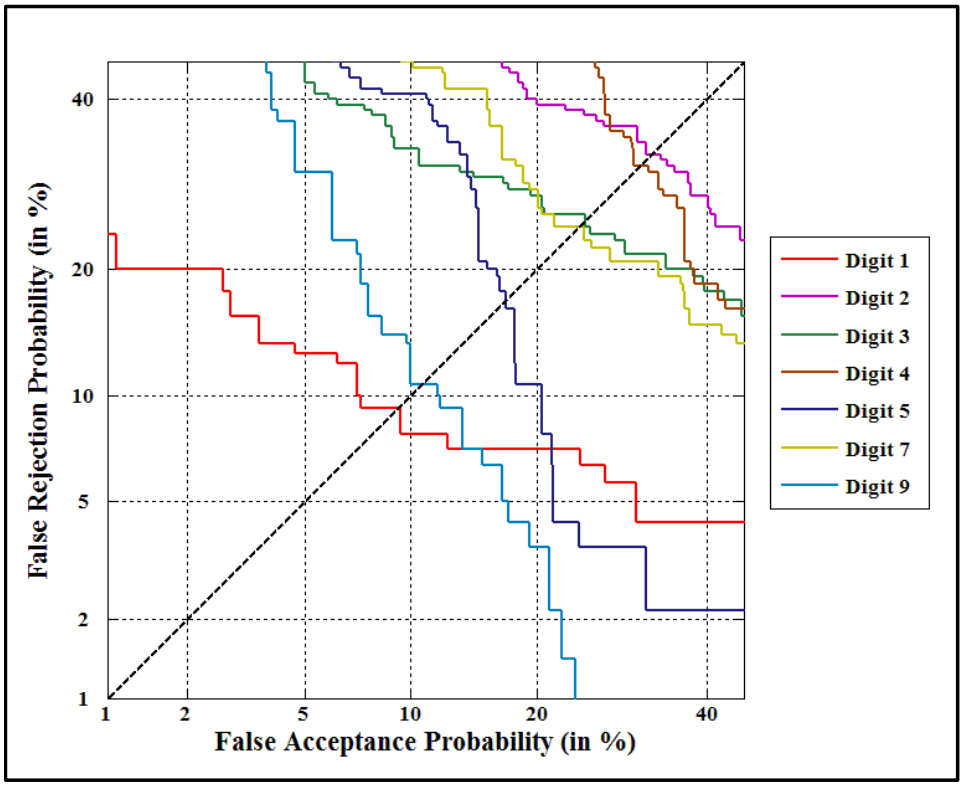

Figure 3.3 The DET Plot for Threshold Estimation using Equal Error Rate criteria for digit models of Spkr-0047

threshold selection are Equal Error Rate (EER) and Weighted Error Rate (WER) [169]. The threshold for the digit classifiers in these experiments is selected based on Equal Error Rate (EER) criterion (on tune/development dataset) which assumes that number of false acceptances and false rejections are equal. The EER for a classifier can be represented in a Detection Error Trade-off (DET) Curve [8] that describes the FRR and FAR in log scales. For example, fig. 3.3 shows the DET plots for the digit models of Spkr-0047. Based on the thresholds selected on development dataset, the performance of the test dataset is evaluated. The next section presents the baseline results for the text-dependent speaker verification tests performed on development and test dataset.

Performance evaluation methods are integral for analysing a verification system and so it is important to select an appropriate reference or baseline system for comparison. The false acceptance and false rejection rates are used to describe the performance of a system. The results for the verification tests can be presented separately for each tune (or test) datasets or the pooled results for all tune (or test) datasets. The error rates in this analysis are presented using the pooled results - as each separate test dataset has a low number of trails that can make the individual results less meaningful. The pooling of results helps in collating the classifier results for each test and then derives the performing characteristics. The classifier results for the tests performed on four train datasets are pooled together to determine the final performance characteristics of proposed fusion. Table 3.3 presents the mean FAR and FRR values for tune and test datasets of eleven randomly selected male speakers from CSLU database. The difference between error rates for tune and test datasets is because of the mismatch conditions or session variability.

The empirical evaluation of proposed architecture performs fusion of different digit classifiers at the decision level. The fusion performance, however, is dependent on the base classifier's performance. Fusing experts of similar and different performances has a

Table 3.3 Verification Performance for Tune and Test Datasets of digits from *SET-1*

SET-1	**Development Datasets**	**Test Datasets**
False Rejection Rate	$0.248^{\pm 0.094}$	$0.253^{\pm 0.095}$
False Acceptance Rate	$0.248^{\pm 0.094}$	$0.247^{\pm 0.095}$

significant effect on the fusion. The empirical evidences of these scenarios have been presented in [30] but lacks the theoretical explanation. The fusion of two classifier decisions may not achieve any improvement in accuracy when the decisions of both classifiers are to either accept or reject the claim. When one system systematically outperforms the other under all conditions, the performance of fusion is similar to that of the best system [30]. Therefore, it is significant to analyse the variations in classifier performances and its effects on fusion.

For a text-dependent system, the error rates obtained are speaker dependent. Even if the vocabulary (texts) used for all speakers is the same, the results for each speaker can be different. It is important to consider the individual speaker performance while tuning the parameters required for fusion, which in turn affects the total fusion performance. Table 3.3 presents the mean verification error rates for the tests performed on digits (1,2,3,4,5,7 and 9) from 11 speakers (i.e., the tests are performed each time choosing one speaker as genuine and the other 10 speakers as impostors). Whereas, the DET curves (in fig. 3.4(a)) provides the comparison between different speakers for tests performed on the same vocabulary. It can be observed that the error rates for base classifiers vary across different speakers. For example, Spkr-0074 has relatively high performance and Spkr-0241 has the worst performance compared to other speakers. The effect of individual speaker's performance on fusion is determined (in the subsequent sections) by analysing the error rates for the speakers with

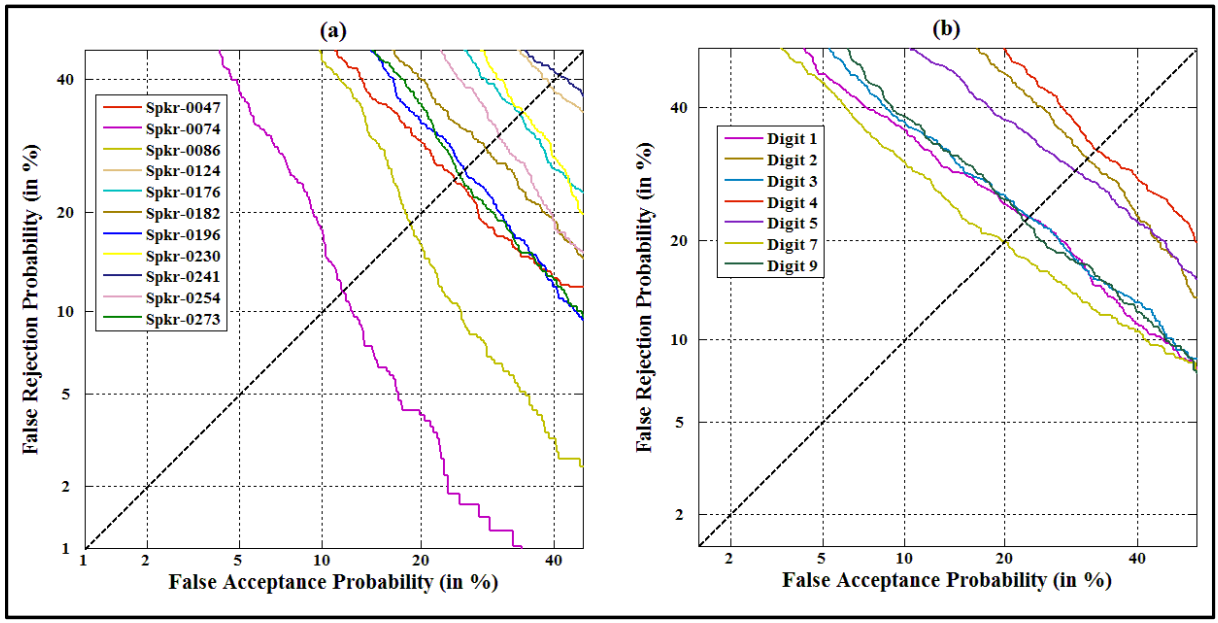

Figure 3.4 DET Plots for text-dependent speaker verification performance of (a) speaker dependent HMM models (b) digit dependent HMM models

good (e.g., Spkr-0074), average (e.g., Spkr-0047) and worse (e.g., Spkr-0241) performances.

The text-dependent speaker verification performance also depends on the vocabulary used for training and testing. The vocabulary in these experiments is isolated digits. Figure 3.4(b) presents the DET curves for each isolated digit with utterances from eleven speakers. It can be noted that models for digits *seven, nine, one* and *three* perform reasonably well compared to other digits. In case of statistically independent decisions, the combination of digit classifiers with lower error rates results in better fusion performance. The choice of the vocabulary used for verification needs consideration in selecting the best set of classifiers with optimal fusion performance. The analysis for determining the set of classifiers with the best performance is done in the chapter 6. The error rates for isolated digit models (irrespective of the baseline performance) are used for validation of the developed expressions for verification errors.

3.5.2.2 *SET-2*

In addition to the dependencies on speaker-specific and digit information, the size of the datasets also helps in determining the efficiency of the fusion scheme. In order to perform the evaluation with greater precision, the size of the test dataset is to be increased. This is achieved by using voice conversion techniques to modify speaker characteristics based on the training data. Impostor testing is performed on casual impostors (speakers other than client). For verification of real impostors, the voice conversion techniques can be employed to transform impostor utterances into client speaker utterances. The conversion is done using the transformation function that replaces the physical characteristics of the voice without altering the message contained in the speech [178]. This dissertation uses the Vocal Tract Length Normalization (VTLN) [179] method in Voice Conversion Matlab® Toolkit [180].

The VTLN approach can compensate for the effect of speaker-dependent vocal tract lengths by warping the frequency axis of the phase and magnitude spectrum. This approach is mainly based on the pitch-synchronous paradigm that can be applied to the frequency or time spectrum. An automatic phonetic class segmentation and mapping approach based on dynamic frequency wrapping is presented in [181] for voice conversion with training data containing different utterances of both source and target speaker. These approaches are used to estimate the parameters of class-dependent VTLN warping functions [181]. The wrapping factor (alpha) and fundamental frequency ratio (rho) parameters are estimated on the training

Table 3.4 System Properties for Voice Conversion using Vocal Tract Length Normalization (VTLN) (The derivations for the parameters and PSOLA is given in section A.1)

	FD-VTLN	**TD-VTLN**
Conversion type	Text-dependent	
Source/target language	English/English	
Alignment Technique	Dynamic time warping	
Warping function Free parameters	piece-wise linear, two segments (α, ρ)	piece-wise linear, two segments $(\overline{w}_1, \rho)$
Acoustic synthesis	frequency domain PSOLA	time domain PSOLA

data for the VTLN-based voice conversion [182]. The default wrapping function used in [180] for both frequency-domain VTLN and time domain VTLN [183] is the symmetric piece-wise function. The detailed explanation of steps for FD-VTLN and TD-VTLN is provided in Appendix A.1. The system properties used for training and estimation of parameters for FD-VTLN and TD-VTLN based voice conversion are given in the table 3.4. The equations for the wrapping factors are also presented in Appendix A.1.

The speech conversion technique employed in this work use FD-VTLN and TD-VTLN approaches. The parameters for VTLN conversion are estimated using the training data (source speaker data) for each digit from a speaker. The estimation and conversion is performed using the Matlab® files '*getwrapingFactor*' and '*vtlnBasedVc*' [180] respectively. The source and target speakers are same for the case of client speech conversion. The data for real impostor testing are obtained by using this conversion method with source speaker data from client and target speaker data from an impostor. The estimated parameters play an important role when assessing voice identities as the values only within a certain range results natural sounding voices [182]. For example, setting v = (1; 1) does not change the voice at all and thus result in the maximum naturalness when distortions by the analysis system can be neglected. On the other hand, extreme values as $\rho \to \infty$ produces artificial or even

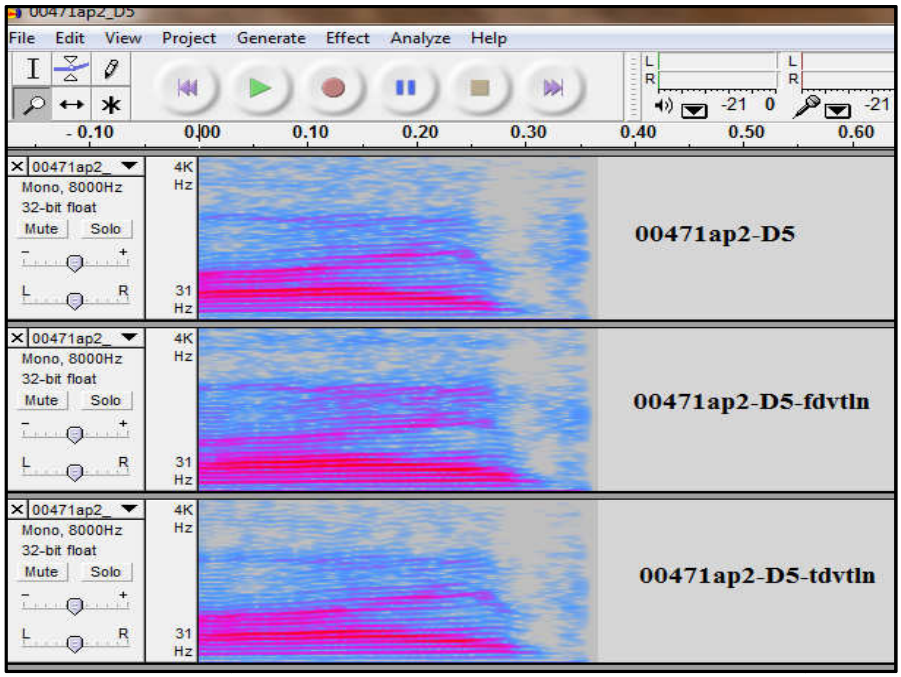

Figure 3.5 Original and converted speech spectral waveforms for digit *'five'*

unrecognizable voices. The estimation and conversion steps are repeated for multiple sets of source and target speech data to reduce the effects of variability in the transformed speech datasets. The parameters used for conversion, here, are trained on source and target data from the same speaker. The parameter estimation is performed independently for each digit of a speaker. For each sample in tune and test dataset, a transformed sample is generated. Depending on the train data, the conversion technique might be good in transforming certain digits more accurately than others. Figure 3.5 presents an example using the spectral representations for FD-VTLN and TD-VTLN converted speech waveforms for single utterance *'digit 5'*.

The protocol similar to the design in section 3.5.2.1 is used to divide the converted data into train, tune and test datasets. Speaker specific digit dependent HMM models are trained for digits from 1 to 9 using 14 client utterances. Each tune dataset has 500 client and 1000 impostor utterances for fine-tuning the parameters, such as threshold estimation and number of base classifier. The test dataset, which is disjoint to the corresponding train and tune datasets, has 1000 client and 10,000 impostor utterances for determining the base classifier performance. The speech data from *SET-1* in addition with the converted data for the digits (1-9) is referred as *SET-2*. Table 3.5 presents the verification error rates for tune and test datasets from CSLU database for *SET-2*. The results presented are for speakers with good (Spkr-0074), average (Spkr-0047) and worse (Spkr-0241) performances.

Table 3.5 Speaker Verification mean error rates for development and test datasets of three speakers from *SET-2*

SET-2	Spkr-0074		Spkr-0047		Spkr-0241	
	FRR	FAR	FRR	FAR	FRR	FAR
Development Datasets	$0.085^{\pm 0.031}$	$0.085^{\pm 0.031}$	$0.249^{\pm 0.068}$	$0.249^{\pm 0.068}$	$0.352^{\pm 0.051}$	$0.352^{\pm 0.051}$
Test Datasets	$0.084^{\pm 0.032}$	$0.084^{\pm 0.032}$	$0.237^{\pm 0.077}$	$0.235^{\pm 0.077}$	$0.355^{\pm 0.050}$	$0.353^{\pm 0.050}$

The fusion performance testing on *SET-2* is analysed in the next chapters for these three speakers. The base classifier performances are usually determined by selecting an optimal threshold on tune dataset for each digit. This selection can be based on either Weighted Error Rate (WER) or Equal Error Rate (EER) [169]. The change in threshold criteria results in different base classifier errors thereby affecting the fusion performance. For this work, results are presented for three operating points that are used in biometric authentication:

- EER for each digit classifier [184]
- Equal FRR for all digit classifiers
- Equal FAR for all digit classifiers [185]

Table 3.6 shows the mean error rates for the tune and test datasets based on the above three criteria. The total error rate values are lower for verification tests performed with Equal error rate (EER) threshold criterion. The next chapter analyses the effect of this difference on fusion performance.

3.5.2.3 SET-*3*

The proposed fusion considers sequential combination of multiple instances and sequential fusion of multiple samples. An instance, here, refers to a digit model and multiple decisions from different digit models are combined using *'AND or OR fusion'* scheme. However, instead of fusing decisions from different digit models, decisions from multiple models for a single digit can be fused. This method of fusion further reduces the vocabulary

Table 3.6 Speaker Verification mean error rates for development and test datasets for three threshold criteria from *SET-2*

SET-2	Equal FAR		Equal FRR		Equal Error Rate	
	FRR	FAR	FRR	FAR	FRR	FAR
Development Datasets	$0.233^{\pm 0.130}$	$0.258^{\pm 0.108}$	$0.255^{\pm 0.106}$	$0.220^{\pm 0.132}$	$0.234^{\pm 0.112}$	$0.234^{\pm 0.112}$
Test Datasets	$0.217^{\pm 0.129}$	$0.265^{\pm 0.113}$	$0.260^{\pm 0.111}$	$0.218^{\pm 0.132}$	$0.234^{\pm 0.112}$	$0.234^{\pm 0.112}$

and length of the utterance (from multiple digits to a single digit) used for verification. The multiple models for a digit can be obtained using different feature extraction/classification algorithms or different training models. The latter approach is employed in this dissertation to investigate the proposed architecture for fusion of different types of classifiers (multiple instances/multiple samples/multiple training models).

The performance of verification system is improved by combining individual classifiers that are both accurate and make errors on a different set of data utterances [186]. The most commonly used method for construction of such classifiers is training data alteration. A set of such classifiers can be generated by training each individual classifier on different subsets of data (e.g. boosting, bagging and k-fold cross-validation). The other advantage of training classifiers with different subsets of training data is that the classifiers obtained are avoided to be highly correlated. Bagging is one of the most commonly known data sampling technique and this approach is used here for creating multiple digit models based on training partitioning.

Bagging technique is the popular way of manipulating the training set [187] where for each run, bagging presents the classifiers with a training set that consists of samples selected randomly with replacement from the original training set. This technique is employed to create multiple models for a digit using the AVICAR database. Five models are created for each digit with each model biased for one of the five-noise conditions (IDL, 35U, 35D, 55U and 55D). The database has 70 utterances for each digit, which is subdivided into train, tune and test datasets. The results obtained from these datasets with limited data may not be

reliable. Therefore, a voice conversion technique [180] is used for obtaining additional client data for each digit. The dataset with original speech data from AVICAR and the converted speech data is referred to as *SET-3*. The protocol explained in section 3.5.2.1 is used for designing the verification experiments using *SET-3*.

A set of 30 utterances (from five noise conditions) is used as training set for creating multiple digit models. The impostor data is obtained from 30 random speakers. The remaining client data is divided into development/tune and test datasets. The speech data in tune set with 500 client utterances and 1000 impostor utterances (equally divided for each noise condition) are used for determining the threshold using Equal Error Rate criteria. For each speaker, the client testing is done on 1000 (200 for each noise condition) utterances whereas impostor testing is done on 10,000 (2000 for each noise condition) utterances. Figure 3.6 presents the DET curves for verification performance for speaker HM3 in *SET-3*. The figure also presents the curves for each individual noise condition. It can be noted that the biased models with IDL condition and 35D has better and worse performance respectively compared to other noise conditions. The use of these biased models for evaluation of sequential *'AND'* and sequential *'OR'* fusion schemes is discussed in the next chapter.

Once the base classifiers are evaluated on the tune and test datasets, the error rates for

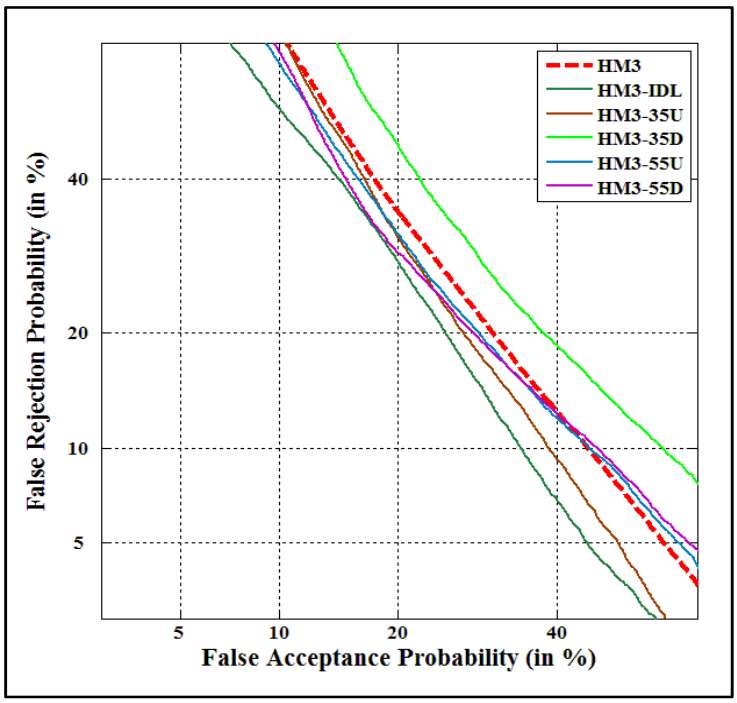

Figure 3.6 The DET Plot for the combined and individual verification performances of five noise conditions (IDL, 35U, 35D, 55U and 55D) for speaker-HM3 in *SET-3*

the fusion of multiple classifiers are calculated theoretically for multi-instance, multi-sample and proposed methods using the developed expressions (2.5 - 2.9). These ideal error rates are then compared to the experimentally obtained error rates. The results for these validations and error rate comparisons are presented in the next chapter.

3.6 Chapter Summary and Conclusion

This chapter provides a basic overview on the classification of speaker recognition systems. The architecture of speaker verification system is detailed with an explanation on feature extraction and modelling techniques specific for text-dependent speaker verification using Hidden Markov Models. The issues of text-dependent speaker verification that should be addressed for an efficient design are also discussed. For a given speaker, the most significant factor that affects performance is the quality and quantity of the training speech. Secondary to these are the environmental factors such as background model, channel variability and threshold estimation criteria. The intraspeaker variations for a speaker should also be modelled while designing the train set used for creating a speaker's model. During testing, the text dependent systems may prompt the speaker for a random subset from the enrolment vocabulary. In addition to these issues, the other factors such as background model design, threshold estimation and protection against spoofing attacks are also discussed.

These issues of text-dependent systems are taken into consideration while designing the HMM based speaker verification system used for empirical evaluation in this dissertation. Text-dependent speaker specific models are trained on limited vocabulary (digit 1-9). In addition, digit-dependent background models are created for appropriate likelihood estimation. The speech data used for these models are from the two databases - CSLU (telephone speech) and AVICAR (microphone speech). The overview of these two databases and protocols used for performance evaluation of the proposed fusion are also detailed. The base classifier performances for the three sets *SET-1, SET-2* and *SET-3* are also presented. The next chapter presents the experimental results for the datasets where the decisions from different text-dependent speaker models are considered statistically independent.

Chapter 4
Empirical Evaluation of Multibiometric Fusion for Text-Dependent Speaker Verification

4.1 Introduction

In the previous chapter, the modelling techniques and challenges for text-dependent speaker verification are discussed. The performance of speaker verification has been shown to improve for fusion techniques where multiple sources of information are considered independent [28, 29, 47]. Sanderson and Paliwal [29] combined the normalized scores from different feature extraction methods, i.e., the approaches for extraction of Mel Frequency Cepstral Coefficients (MFCC) and Maximum Auto-Correlation Values (MACV), for the same utterance to improve the verification performance. Cheung et al. [26] proposed the fusion of scores from multiple speech samples based on their score distribution and prior score statistics. Ramli et al. [47] studied the combination of information from multiple instances of three verbal models (*zero, seven* and *eight*) and multiple biometric modals (speech and face subsystems). The scores for verbal samples are fused using sum-rule and weighted sum-rule fusion resulting in EER of 2.03% and 4.32%, respectively. In this chapter, multi-instance and multi-sample fusion schemes are evaluated for improvement in speaker verification performance. The speech data are evaluated using HMM based verbal models for combination at decision level assuming independence between the decisions from verbal models.

The sequential fusion method is evaluated for text-dependent speaker verification for variations in the amount of available data and the order of presentation (section 4.2). In the subsequent sections, the performance is evaluated for multi-instance fusion, multi-sample fusion and proposed multi-instance and multi-sample fusion schemes for performance improvement. In section 4.6, the architecture performance is compared for combination of multiple models with other biometrics. The statistical analysis for the comparison between experimental and theoretical error rates for the proposed fusion is presented in section 4.7. This section also explains the effect of base classifiers on the difference between the ideal/theoretical and experimental error rates.

4.2 Sequential Decision Fusion

OVERVIEW: This section presents a brief introduction on the sequential fusion approach and its application in pattern recognition application. The sequential approach improves user convenience during acquisition and provides higher throughput during processing. The serial processing approach effectively reduces the processing time as a decision can be made without waiting for the outputs of all the biometric subsystems (instances/samples). In this approach, the user is provided with ability to choose the order in which available inputs can be processed. This section evaluates the impact of the sequential method on the trade-off between performance and amount of data required for making a reliable decision. The evaluation results are also presented for the difference in order of processing multiple instances in the context of Text-Dependent Speaker Verification.

The sequential fusion approach has been widely used in biometric literature. Jacek et al. [188] proposed a multi-frame multi-expert system that uses the sequential fusion of scores obtained on successive video frames of a user's face to reduce the error rate. Poh et al. [189] suggested that the sequential fusion approach can be used to minimize the cost, i.e., a sequential fusion algorithm can be used to match scores until a desired confidence is reached, or until all the match scores are exhausted, before obtaining the final combined score. The sequential approach is also beneficial in making *reliable decisions* using *smaller blocks of data* as they become available rather than wait for the entire utterance. For example, in [93], the SPRT approach is used to demonstrate that accurate verification decisions are obtained after only 2-10 seconds of evaluation data where usually 100 seconds are needed. Surendran [21] also proposed the use of sequential method for a systematic trade-off between the performance of the system and the amount of data needed to make a decision. The analysis used two different thresholds (upper & lower) to determine the trade-off between average error and average sample number (ASN) needed for making the decision. The experimental evaluation has shown that ASN of 2.85 and 4.71 is required for true speakers and impostors respectively for an average error of 3.1%. The reason for high impostor ASN is explained to be the liberal lower threshold. This analysis, however, does not consider the *order in which the samples are combined* which also has an impact on system performance [131].

The proposed sequential decision fusion architecture (section 2.5) is theoretically shown to control the trade-off between verification error rates (FRR & FAR). Before

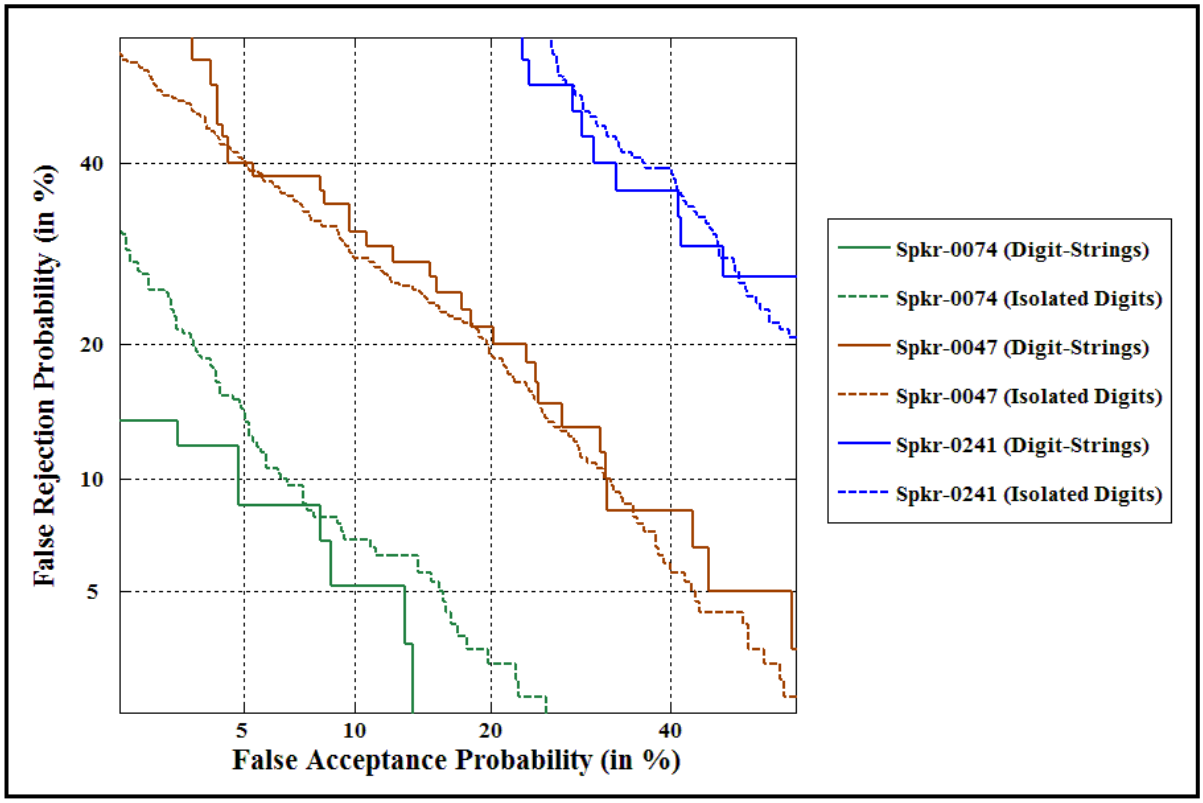

Figure 4.1 DET plot for the baseline performances of Digit-Strings and Isolated Digits of three speakers (Spkr-0074, Spkr-0047 & Spkr-0241)

proceeding to empirically evaluate this architecture, the impact of sequential fusion method on the trade-off between performance and number of decisions required for reliable decision is determined. The other factor for consideration is the sequence in which decisions are combined. The experimental results are presented for speaker verification test using six digit-strings (sequence of 5-digits) for three speakers in CSLU database.

The data for each digit-string is divided into a train set of 12 utterances (one utterance from each one of the sessions) and a test dataset (30 client utterances and 300 impostor (30*10) utterances from 10 different speakers). Each utterance in the dataset is segmented manually into isolated digits using Audacity Software [177]. The threshold criterion for digit-strings and isolated digits are the Equal Error Rate (EER) and Equal False Rejection Rate (FRR) respectively. Figure 4.1 plots the performance of Digit-Strings and Isolated Digit models for the three speakers (good (Spkr-0074), average (Spkr-0047) and worse (Spkr-0241)). The baseline performances for both the digit-strings and isolated models are observed to be similar for a speaker.

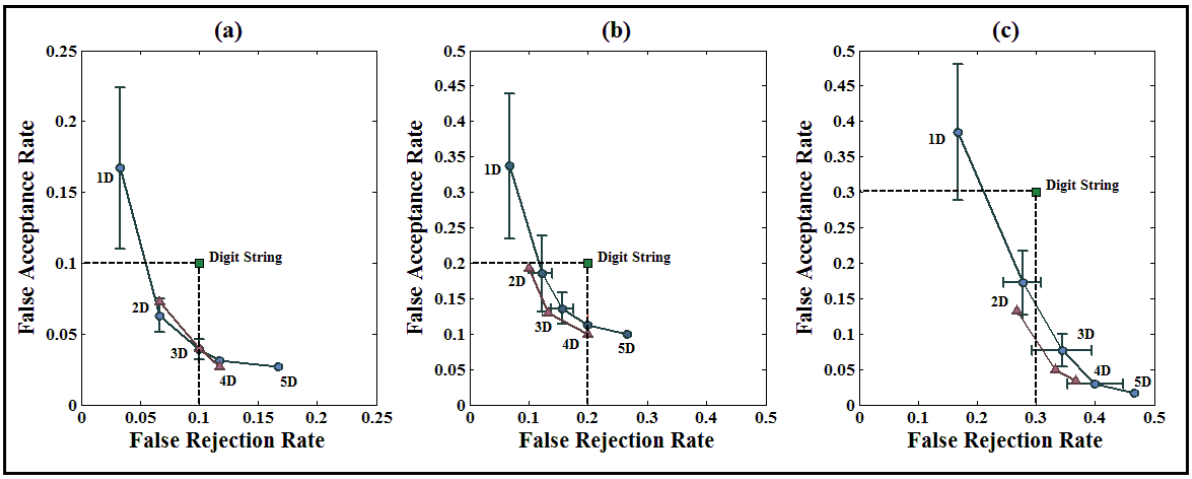

Figure 4.2 Error rates for the digit-string (2-8-3-7-6) and sequential fusion of Isolated Digits for (a) Spkr-0074, (b) Spkr-0047 and (c) Spkr-0241 (The points in brown represent the error rates for fusion of two (2D), three (3D) and four digit (4D) decisions in the order similar to the digit-string. The points in blue are the mean error rates for sequential fusion of digits in all possible combination sequences. The point in green represents the FRR and FAR for the digit-string)

The error rates for the fusion of isolated digits are compared to the corresponding digit-string models. Figure 4.2 presents the comparison of false rejection and false acceptance rates for digit-string and sequential fusion of decisions from isolated digits (2, 8, 3, 7 and 6). The points on blue line represent the mean error rates (with standard deviations) for fusion of two (2D) to five digit (5D) decisions for the digit-string. The points on brown line represent the error rates for sequential fusion of digits (2D, 3D and 4D) in the order same as the digit-string. Lower error rates are obtained from the fusion of digits in the order same as digit-string (i.e., on brown line) rather than random fusion of digits (i.e., blue line points). Further, the error rates for sequential fusion of isolated digits (two and three digits from various digit-strings) are observed to be lower than that of corresponding digit-string (sequence of 5 digits). Although isolated digit models have high FAR, the false accepts decreases with an increase in number of fused decisions. The FRR for the sequential fusion of available digit models, however, is high compared to the digit-string model.

The improvement in performance for fusion of isolated models can soon reach saturation. The error rates for sequential fusion of two and three isolated digit models (from a digit string) are observed to have better performance (than digit string model) whereas the combination of four and five digits may not often result in performance improvement. The

improvement in fusion performance is shown for the sequential fusion of decisions from half the number of the individual digits in the digit-string. From fig. 4.2, it is also demonstrated that the sequential fusion approach has better control over trade-off between performance and number of decisions used for fusion. It is also shown that, in general, the sequential fusion of decision in the order similar to digit-string has better performance compared to the mean values of all possible isolated digit combinations. With the increase in number of decisions, the number of false rejects increases, which is reduced by allowing the user with another attempt for the rejected digit. The approach of multiple presentations could be applicable when digit models have low FAR (and high FRR) as the fusion of attempts/samples results in increasing the FAR (while reducing the FRR). Therefore, the trade-off between ***verification error rates*** and ***amount of time taken to reach a decision*** is better controlled by varying the number of digits combined in the sequence along with the number of repetitions at each digit.

The next section presents the extended evaluation of multi-instance architecture that employs sequential fusion of decisions from multiple instances (digits in this scenario). The integration of multi-instance fusion with a multi-sample fusion scheme enables better control over the trade-off between the verification error rates (false rejection rate and false acceptance rate). This architecture is applicable to telephone and internet shopping applications where remote authentication is performed using speaker verification. It is important in these applications to serve both security and user convenience requirements achieved by deciding on the number of attempts (samples) at each decision stage and the number of decision stages (instances) used for verification.

An ***instance*** in the context of text-dependent speaker verification architecture refers to the ***text or digits*** that form the decision stages. A ***sample*** here represents any ***single utterance*** of a digit from a speaker. If a sample is rejected at a decision stage, the next sample is selected in random from the remaining utterances or an adaptation of rejected sample to the claimed model.

4.3 Fusion of Multiple Instances

OVERVIEW: Fusion of multiple instances within a modality for biometric speaker verification performance improvement has received considerable attention [47]. For greater accuracy, the decisions from multiple instances are combined sequentially and the fusion is analytically shown to improve performance under the assumption of statistical independence

(Section 2.7.1). This section provides the empirical evaluation of equations developed for verification error rates considering the architecture for text dependent speaker verification using Hidden Markov Model (HMM) based digit dependent speaker models. The tuning of parameters, n classifiers/instances, is investigated and the resultant verification error trade-off is evaluated on individual digits. The multi-instance fusion scheme is evaluated for the variations in threshold criteria, datasets and models used for speaker verification.

The architecture of a multi-instance scheme for text-dependent speaker verification is shown in fig. 4.3. There is a sequential chain of text-dependent digit classifiers with each classifier verifying an input utterance/sample for the particular digit. The classifier 'HMM_i' is modelled using the training data of instance '*i*'. Whenever classifier HMM_i (*i*=1, 2, 3...*n*) accepts the input data, the control is given to acquire input for the next classifier in the sequence, HMM_{i+1}. In a sequential decision framework, a speaker is to be accepted by all instances/classifiers in the sequence. Acceptance is thus logical *'AND'* for multiple instances. If the speaker is rejected at any decision stage, the sequence terminates and thus rejection decisions are logical *'OR'* for multiple instances (section 2.7.1.2).

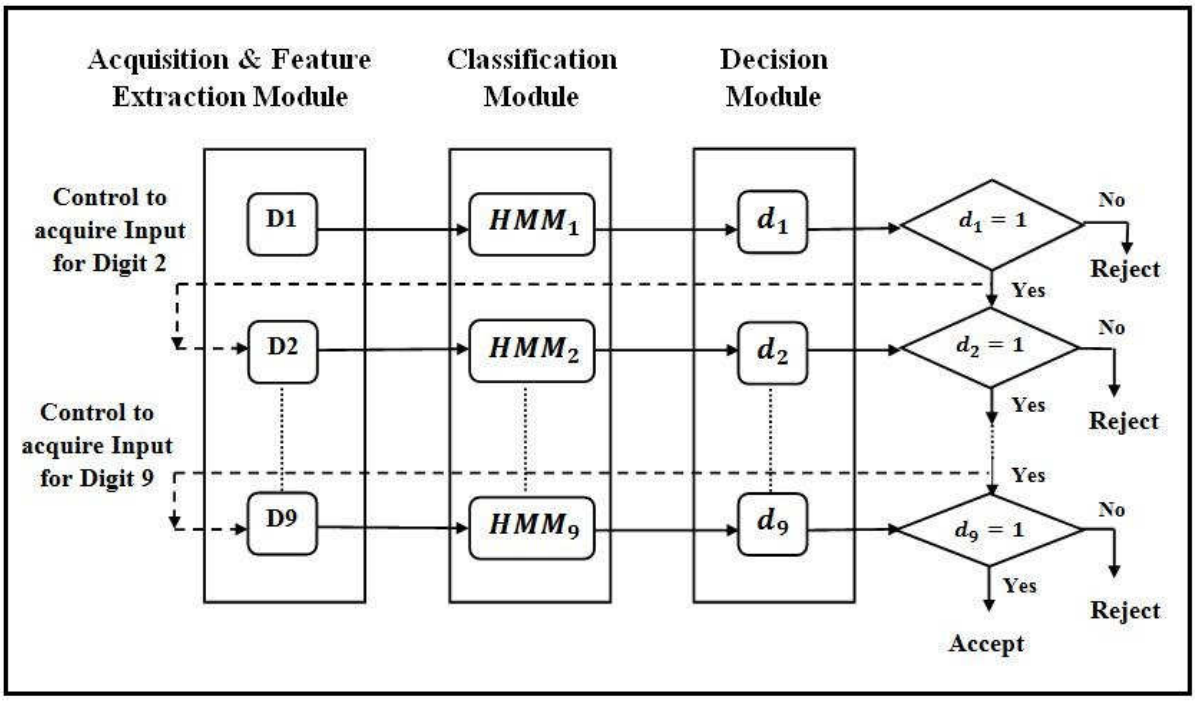

Figure 4.3 The architecture of a sequential multi-instance fusion scheme with '*n*' classifiers

The protocol used for the performance evaluation of multi-instance fusion is described in section 3.5.2. An ***instance***, here, refers to a ***digit*** that is modelled using a HMM to represent a speaker's characteristics. Therefore, in this section, ***multi-instance fusion*** is referred as the ***combination of digits***. The scheme is evaluated for speech data from *SET-1*, *SET-2* and *SET-3*. The results for all possible combinations of *'n'* classifiers are used to obtain reliable estimates of error rates.

The fusion performance of the system is tested by progressively increasing the number of instances/digits used for verification. Figure 4.4 plots the false rejection and false

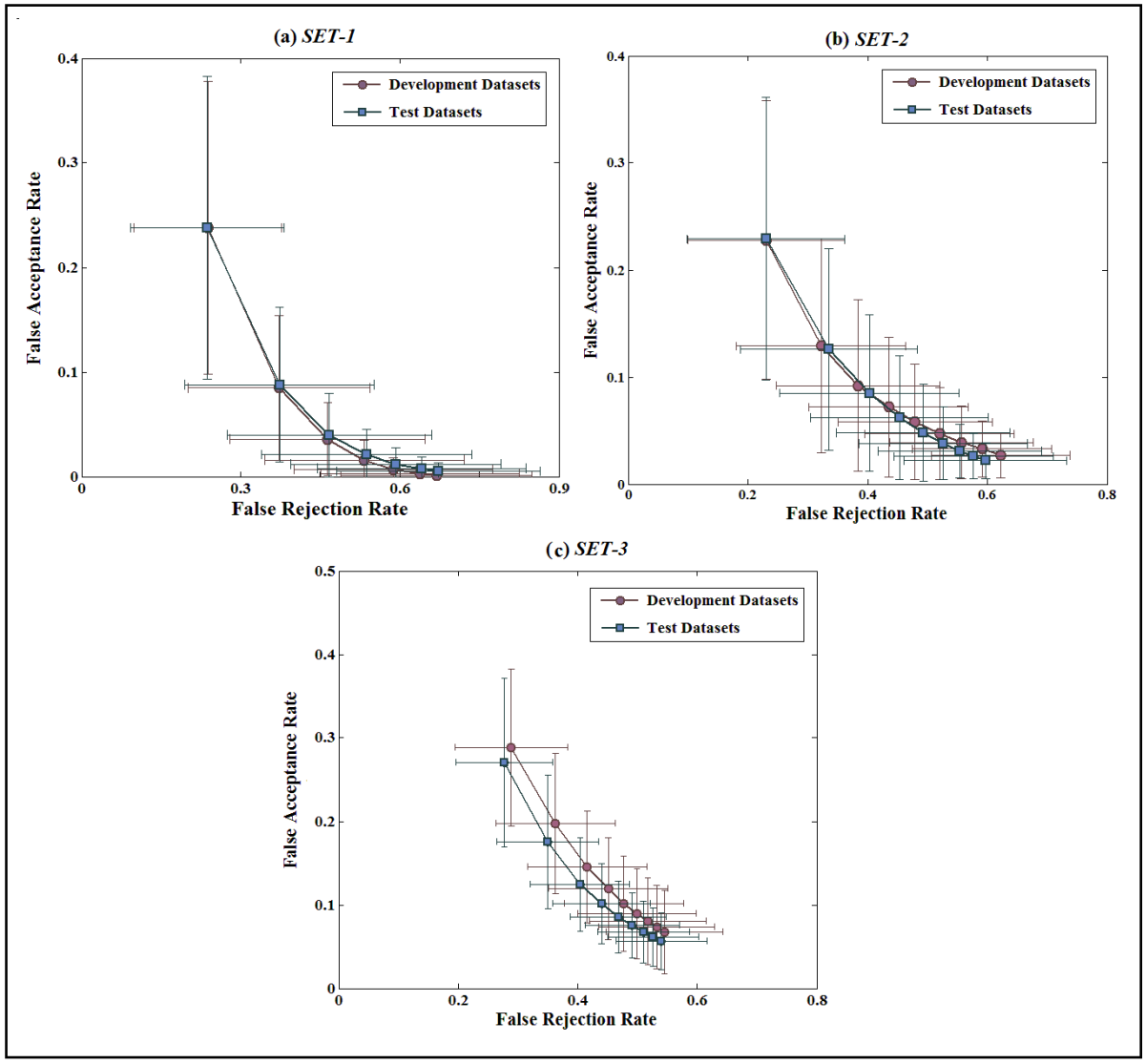

Figure 4.4 Speaker Verification Performance for development and test datasets of (a) *SET-1*, (b) *SET-2* and (c) *SET-3* (Each point on the curve represents FRR & FAR for sequential fusion of digits. The points to the top-left of each curve represents errors for isolated digits (1D) and then increases progressively to 7D for *SET-1* and 9D for *SET-2 & SET-3*)

acceptance rates for the fusion of multiple instances from development/tune and test datasets respectively. The plots are shown for three sets - *SET-1* (fig. 4.4(a)), *SET-2* (fig. 4.4(b)) and *SET-3* (fig. 4.4(c)). The point to the top left in each figure represents the mean error rates with standard deviation for speaker verification on isolated digits (1D). The subsequent points on the curves represent the error rates for each progressive addition of digits used for fusion i.e., second point (2D) gives the mean FRR and FAR with standard deviation for the tests on two-digit combinations, third point (3D) is the mean error rate for tests on three-digit combination and so on. The last point for *SET-1* represents the fusion errors for seven digits whereas for *SET-2* & *SET-3* the last point represents errors for nine-digit combination. The number of false acceptances is lowered at the cost of an increase in false rejections for multi-instance fusion. The fusion of seven digits results in reducing the false acceptance rate by 23.7% and increases the false rejections by 43% for test datasets of *SET-1*. The number of false acceptances and false rejections for combination of nine digits are 13.6% and 59.7% respectively for *SET-2*.

The multi-instance fusion performance, in general, improves when the decrease in FRR is greater than the increase in FAR. The Total Error Rate, TER (FRR+FAR), for three *SET*s (fig. 4.4) mostly increase the TER for the fusion of digits. For example, the TER decreases from 54.8% to 52.5% for verification based on isolated digits and fusion of two digits respectively. This TER increases with each additional digit and thus for nine-digit fusion, the TER increases to 59.7% (FRR - 54% & FAR - 5.7%). Though error rates in fig. 4.4 represent pooled results for all speakers in the datasets, the analysis when extended to individual speakers has shown the same conclusions.

The error rates from tune dataset can be used to estimate the fusion performance theoretically on test dataset. When the tune and test datasets (disjoint sets) are of equal base error rates, the theoretical error rates for fusion of different digits are also similar when the decisions between the classifiers are considered statistically independent. This may not be the case with experimental error rates. For example, fig. 4.4 shows the mean FRR & FAR errors for tune/development and test datasets that are disjoint with different sizes. Here, the base performances for tune and test datasets are different for *SET-3*. Although the mean error rates are similar for both *SET-1* and *SET-2*, the error rates for the isolated digits are different. The experimental results of the fusion of digits (tune and test datasets) are shown to have mean error rates that are within the range of the standard deviation.

The fusion performance, in general, depends on base classifier performances, either similar or different [190], and if a single classifier outperforms the other, the fusion performance is similar to that of the best classifier [30]. In *SET-2*, the digit models are with similar (equal FRR and equal FAR threshold criteria) and different performances. In table 4.1, the error rates for fusion of digits with similar and different performances are presented for tune datasets of *SET-2*. Better error rates are obtained for the fusion of classifiers/digits with different rather than similar base errors. As fusion performance depends on base performances, the error rates are lower for fusion of classifiers with different rather than similar performances. In addition, the difference in TER (*fusion TER-base TER*) is greater for classifiers with *different performances* compared to classifiers with similar base performances. For this dataset, the fusion of two digits improves the performance by 0.2% compared to base performance. However, performance degrades with an increase in digits (-ve) used for fusion, as the decrease in false accepts here is lower than an increase in false rejects.

Table 4.1 Error rates for multi-instance fusion of digits with **s**imilar and different base performances (positive difference (+) here refers to case where *fusion TER* < *base TER* and negative difference (-) here refers to case where *fusion TER* > *base TER*)

Number of digits	Similar Error Rates (in %)		Different Error Rates (in %)		Difference between base and fusion TER	
	FRR	FAR	FRR	FAR	Similar base performance	Different base performance
1	25.5	25.8	22.3	22.0		
2	35.0	16.1	32.3	11.7	0.2 (+)	0.2 (+)
3	41.2	13.0	39.1	8.6	2.9 (-)	3.4 (-)
4	46.3	11.1	44.4	6.9	6.1 (-)	7.1 (-)
5	50.7	9.8	49.0	5.7	9.2 (-)	10.5 (-)
6	54.5	8.8	53.1	4.9	12.0 (-)	13.8 (-)
7	57.8	8.0	56.8	4.2	14.5 (-)	16.8 (-)
8	60.8	7.4	60.1	3.7	16.8 (-)	19.5 (-)
9	63.5	6.8	63.1	3.2	18.9 (-)	22.1 (-)

As the base error rates are different between speakers, the improvement in performance for fusion of digits is speaker-dependent. With the fusion of multiple instances (digits), the number of *false rejections* increases whereas the *false acceptances* decrease irrespective of the base classifier performances. For example, fig. 4.5 represents the error rates for sequential *'AND fusion'* of digits for three speakers of different base performances (Spkr-0074, Spkr-0047 & Spkr-0241) from *SET-1*. Similar results for rest of the speakers from *SET-1* are presented in the fig. 4.27. For each speaker, the points on curves represent mean error rates (FRR & FAR) with standard deviation for different digit combinations. The increase in FRR and FAR for the fusion approach is directly and indirectly proportional to increase in instances/digits used for fusion. The fusion of two digits results in reducing the FAR by 14.8% (21.6%) and increasing the FRR by 13.2% (15.8%) for the Spkr-0047 (Spkr-0241). The overall performance of fusion is shown to improve when the decrease in false accepts is higher than the increase in false rejects. The overall performance improves for two-digit combination but then the total error rates increases with each additional digit - the combination of seven digits results in 38.8% (21.1%) decrease in FAR whereas 43% (43.3%) increase in FRR for Spkr-0241 (Spkr-0047). The fusion performance is thus shown to be dependent on both ***base performance*** and ***the number of digits*** to be used for fusion.

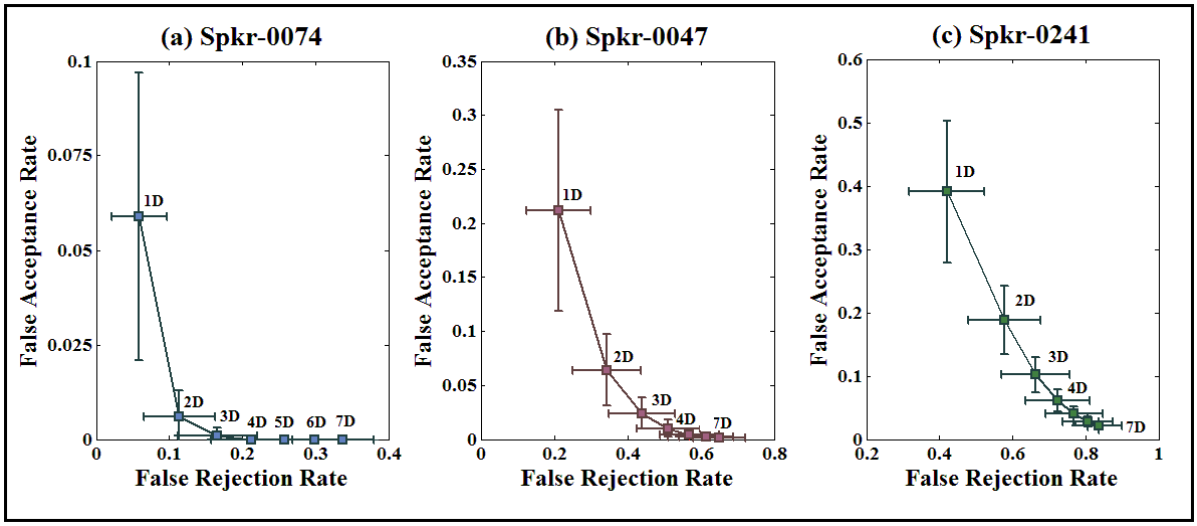

Figure 4.5 Speaker dependent verification error rates for fusion of digits from *SET-1* (a) Spkr-0074, (b) Spkr-0047 and (c) Spkr-0241 (Each point on the curve represents FRR & FAR for sequential fusion of digits. The points to the top-left of each curve are for isolated for isolated digits (1D) and the last point is for the seven digit (7D) combination)

The speaker specific base-classifier error rates, in general, depend on the feature extraction and classification methods used for verification. Apart from these technical aspects, the performance of the base classifiers also depends on the data used for training and verification, dataset size and session variability. For disjoint test datasets, base performances can be either similar or different (fig. 4.4). Fusion of multiple instances (digits) is shown to result in error rates (with standard deviation) that are proportional to the base performance of a speaker across different datasets. The same conclusion is extended for verification tests performed on datasets for individual speakers (fig. 4.28) with data overlap. In addition, different models for the same speaker dataset also lead to differences in base classifier performances. Figure 4.6 (a) and (b) represents the false rejection and false acceptance rates for combination of digits from same test-dataset with different training models (model for each noise condition in *SET-3*) respectively. If the base error rates for each training set are significantly different from each other, then fusion performance is enhanced by selecting the training models with *minimum error rates* (for each digit or all digits). However, for models with relatively similar base performance, the models with low FRR and FAR may not always ensure minimum fusion error rates. For example, in fig.4.6, the use of IDL model for testing results in better base performance compared to others and so the fusion FAR is minimum

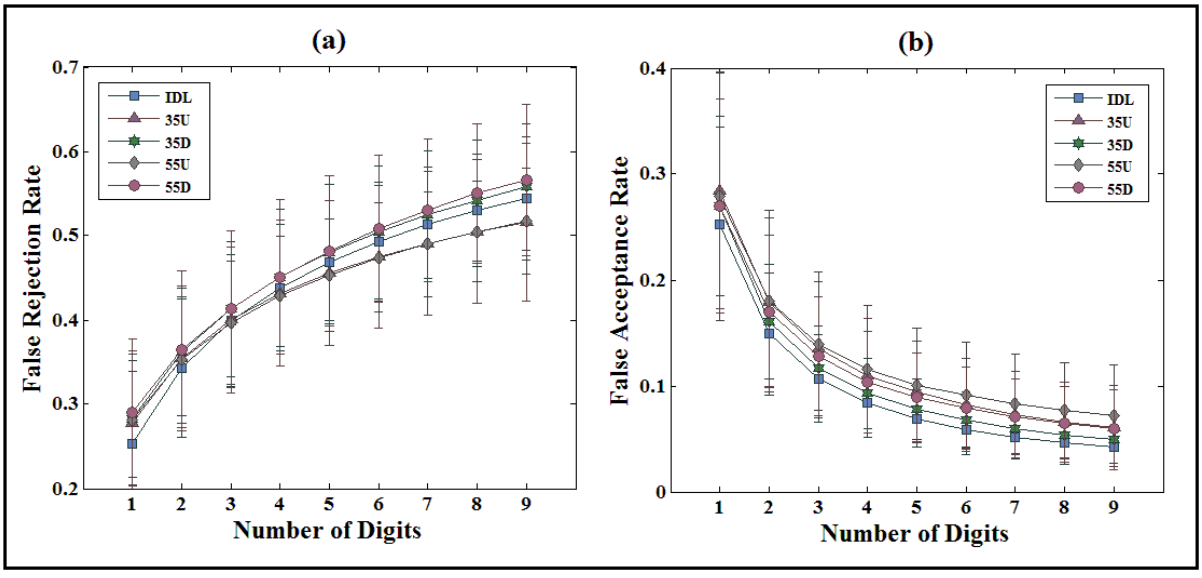

Figure 4.6 Error rates for sequential fusion of digits with different training models for the same dataset (a) False rejection rates and (b) False acceptance rate (The bottom-left point in (a) and top-left point in (b) represents errors for isolated digits then increases progressively to fusion of nine digits)

for this model but the FRR for fusion is lower for the models of samples with windows up (35U & 55U). Therefore, the deciding factors for fusion performance here are the base performance and the complementary information from the decisions of different digits. These factors are discussed in detail in next chapter. Though the analysis here is explained for multiple training models of same dataset, the conclusion can be extended for the use of multiple training models on different datasets (Figure 4.29 presents the fusion error rates for the different datasets verified on different training models for three speakers from *SET-1*. It is shown in this figure that the models with low FRR and FAR does not always ensure minimum fusion error rates. For example, verification tests performed using train set-4 results in base FRR (FAR) of $0.391^{\pm 108}$ ($0.394^{\pm 108}$) and fusion FRR (FAR) of $0.843^{\pm 0.031}$ ($0.020^{\pm 0.003}$) for seven digits. Similarly, the use of the train set-2 models results in base FRR (FAR) of $0.416^{\pm 0.80}$ ($0.418^{\pm 0.81}$) and fusion FRR (FAR) of $0.796^{\pm 0.018}$ ($0.019^{\pm 0.003}$) for seven digits. Although the base performance is better for train set-4 models, the fusion error rates are lower for the combination of seven digits from train set-2).

The other factor that affects the base classifier performance is the choice of the threshold criteria used for verification. Figure 4.7 presents the mean error rates for the combination of digit decisions obtained using three different threshold selection criteria (section 3.5.1). The first criterion is based on adjusting the threshold to obtain equal error rate i.e., equal FAR and FRR, for each digit model. The second criterion is to obtain equal FAR for all the digit models whereas the last criterion is to select the thresholds to obtain equal FRR for all the digit models. The fusion of multiple instances is shown to increase FRR and decrease FAR. But the increase in FRR decreases with an increase in the number of digits combined in sequence i.e., in fig.4.7 the FRR increases by 10.6% when two digits are combined whereas this increase reduces to 4.2% and 2.8% for three and four-digit combinations respectively. Similarly, the decrease in FAR is lowered with an increase in the number of digits in the sequence. For example, the decrease in FAR is 12.6%, 9% and 7.4% for two, three and four digit sequences for the fusion using Equal FAR.

Therefore, the FRR increases and FAR decreases with an increase in instances for multi-instance fusion. This increase in FRR or decrease in FAR is lowered with each progressive addition of an instance. The increase in the number of false rejections can be reduced by allowing fusion of multiple samples at each instance. The difference in fusion performance for different datasets/thresholds/models is due to the base classifier performance and the dependence between the classifier decisions. This difference is observed to be lower

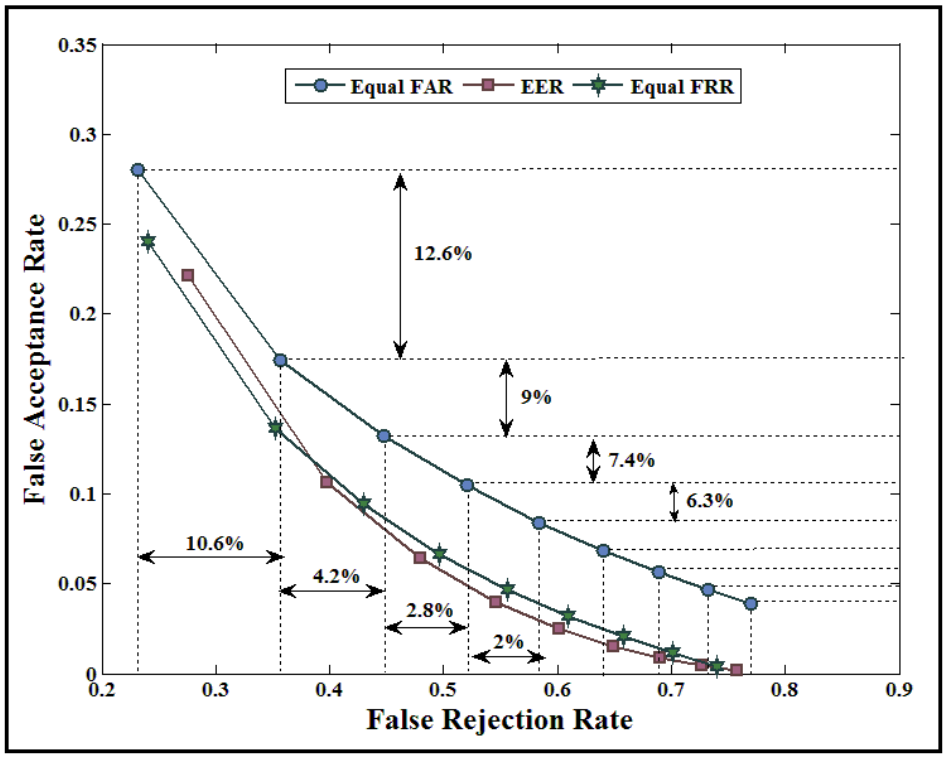

Figure 4.7 Verification error rates for multi-instance fusion of a dataset with different threshold selection criteria for Spkr-0047

when the mean error rates for different threshold criteria are similar. However, the best threshold criterion for verification using proposed architecture is better determined by fusing the decisions from multiple instances where multiple samples are allowed for each instance (section 4.5.4.2). The next section, however, deals with the analysis of multi-sample fusion technique for text-dependent speaker verification.

4.4 Fusion of Multiple Samples

OVERVIEW: Fusion of multiple samples from a single modality has been shown to improve speaker verification performance [26]. *The sequential combination of multiple samples using 'OR Rule' is analytically shown to improve performance (section 2.7) when the decisions from multiple samples are assumed to be statistically independent. The multi-sample fusion scheme is empirically evaluated of considering the architecture for text dependent speaker verification using Hidden Markov Model (HMM) based digit dependent speaker models. The tuning of parameters, m samples of a classifier/instance, reduces the false rejects at the cost of an increase in false accepts. The multi-sample fusion scheme is evaluated for the*

variations in datasets and models used for speaker verification. In addition to these, the fusion performance is evaluated for the changes in the nature of subsequent samples, adaptive or random presentations, used for the verification in case of a rejected sample.

Figure 4.8 presents the architecture of multi-sample speaker verification system (section 2.7.1.2). Considering *'m'* to be the maximum allowed number of repeated samples and $X_i(i=1, 2 \ldots m)$ to be the input test utterances from the speaker, the classifier *HMM* makes a decision to either accept or reject the speaker. For a speaker to be declared genuine, for a particular instance (or spoken text), it is considered sufficient if any one sample (or utterance) presented to the system gets accepted. Acceptance decisions are logical *'OR'* for multiple samples. However, if the speaker is accepted by 'i^{th} sample' ($1<i<m$) then the subsequent samples need not be verified. The speaker is considered to be an impostor when all the '*m*' samples are rejected. Rejection decisions are thus logical *'AND'* for multiple samples. Once the verification test for the sample is rejected, the next sample to be used for fusion significantly changes the resulting performance. The nature of the repetition sample for a rejected sample can either random or adaptive.

- In case of a true speaker, the ***random sample*** is another presentation of the required utterance. The random sample for an impostor is a naive or zero-effect attack where the impostor is trying to be accepted by system without the knowledge of actual speaker's speech characteristics. For experiments presented in this section using random samples, each utterance of a digit in the tune/test dataset is presented as a sample to the speaker-specific model of that digit and repetitive samples are randomly picked from the remaining dataset.

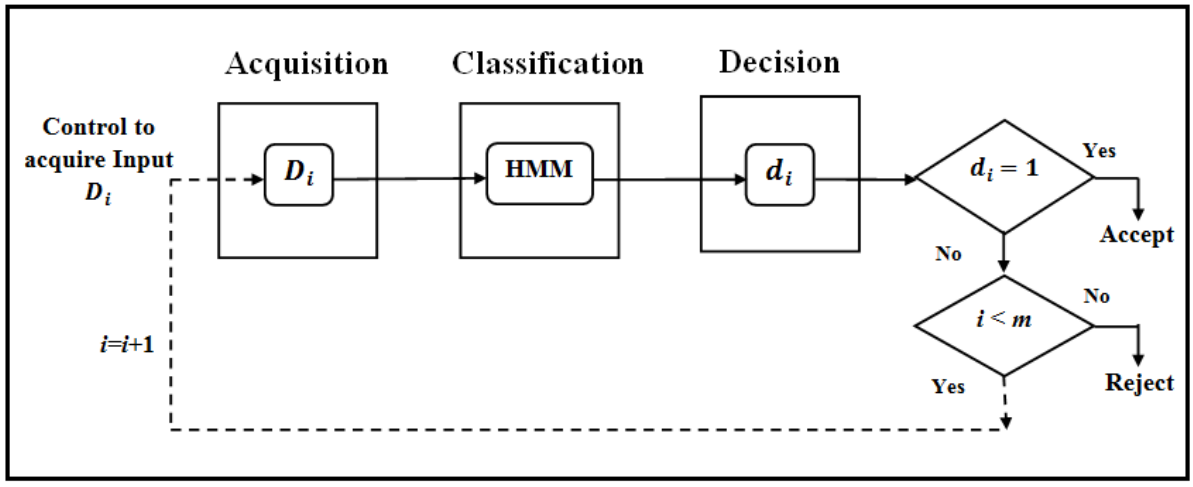

Figure 4.8 The architecture of a multi-sample fusion scheme with *'m'* repetitive samples

- The ***adaptive samples,*** in case of a client or an impostor, are attempts to change his/her characteristics to adapt to the claimant model. An impostor can try to adapt to the claimant model by mimicking the claimant speaker. This could also be achieved using the techniques and/or tools for obtaining transformed impostor utterances from the claimant utterance. With each repeated attempt, an impostor can improve the chance of adapting characteristics similar to claimant model. An adaptive sample, in this dissertation, is obtained using voice conversion techniques. For each utterance of a digit in the tune/test dataset that is rejected, a transformed utterance of the same sample from the conversion technique is used as the repetitive sample.

4.4.1 Random Samples

In this section, the empirical evaluation of multi-sample fusion at the decision level is explained. The protocol used for this performance evaluation is described in section 3.5.2. A ***sample*** for an *instance*, here, refers to an ***utterance*** for a *digit* that is used for speaker verification by a HMM model that represents the characteristics of a speaker. Therefore, in this section, ***multi-sample fusion*** is referred as the ***combination of decisions from multiple utterances of a digit***. The scheme is evaluated for speech data from *SET-1*, *SET-2* and *SET-3*. In this work, the analysis is presented using the pooled results for all test datasets.

The fusion performance of the system is tested by progressively increasing the number of samples used for verification. The overall performances of development/tune and test datasets for the fusion of multiple samples from *SET-1, SET-2* and *SET-3* are presented in fig. 4.9. The first point in figure represents the mean error rates with standard deviation for speaker verification using isolated digits. The subsequent points on the curve represent the error rates for each addition of a repeated sample used for fusion i.e., second point gives the mean FRR and FAR with standard deviation for the tests on two-sample fusion and so on. The number of false rejections decreases and the false acceptances increases for multi-sample fusion. For example, the fusion of two samples or the use of one repeated sample reduces FRR from 23.7% to 7.4% whereas increases FAR from 23.8% to 39.9% for test dataset of *SET-1*. With the increase in samples used for fusion, the FRR decreases further but do not ensure improvement in overall fusion performance because of a greater increase in FAR. The fusion of five samples reduces the false rejection rate by 22.8% and 27.3% for test datasets of *SET-2* and *SET-3* respectively. The increase in false acceptances is higher than the decrease

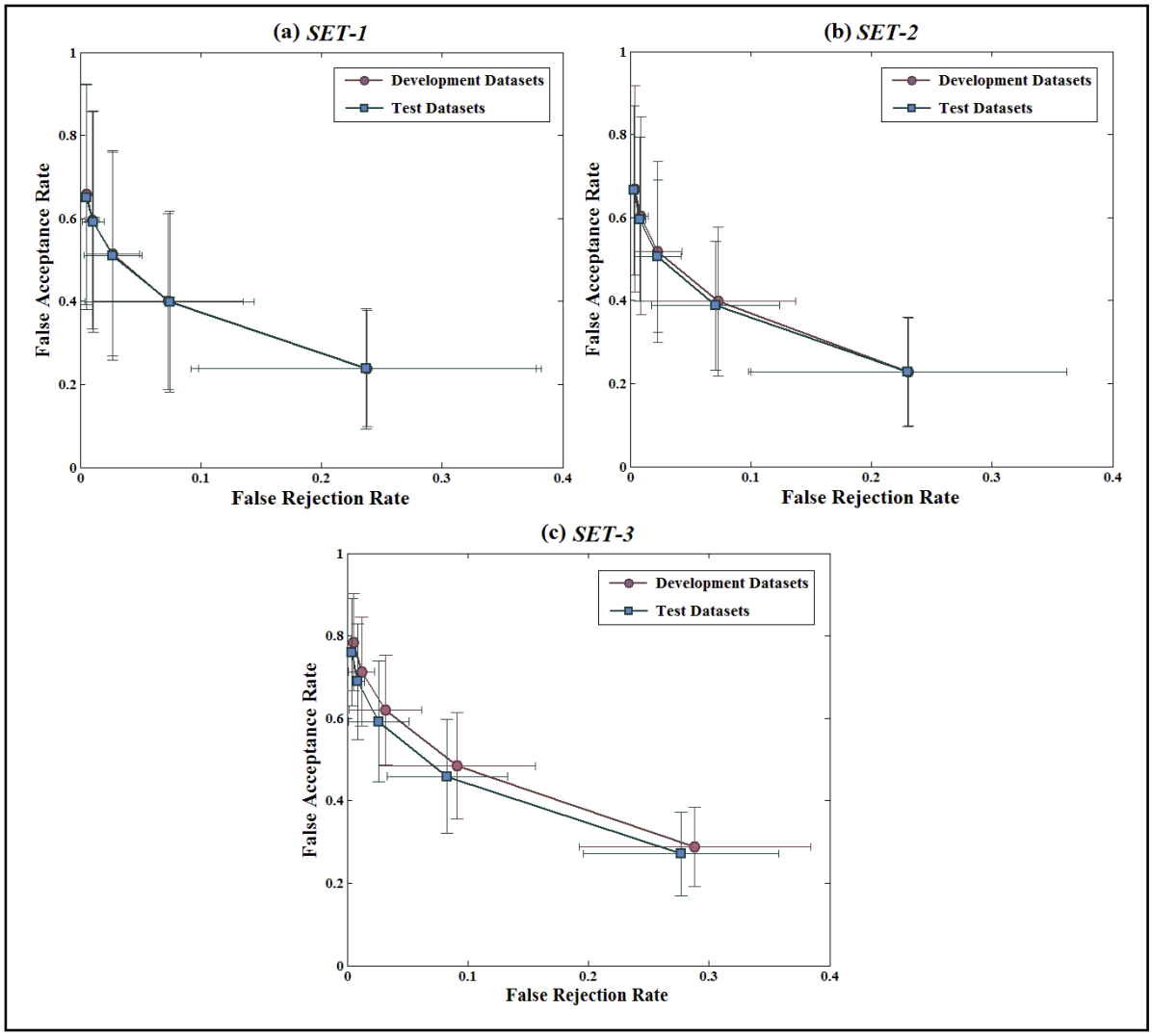

Figure 4.9 Verification error rates for fusion of samples in development and test datasets from (a) *SET-1*, (b) *SET-2* and (c) *SET-3*

in FRR with digits used for fusion, i.e., FAR increases by 43.7% and 48.8% for *SET-2* and *SET-3* respectively. The fig. 4.9 represents the error rates for both the development/tune and test datasets, which are considered to be, disjoint with different sizes. The fusion performance of the two datasets (tune and test) is shown to have mean error rates that are within the range of the standard deviation provided the base classifier performances are similar. When the tune and test datasets (disjoint sets) are of similar base performances, the fusion performance of the test set is estimated using base error rates of tune datasets when the decisions between the samples are considered statistically independent.

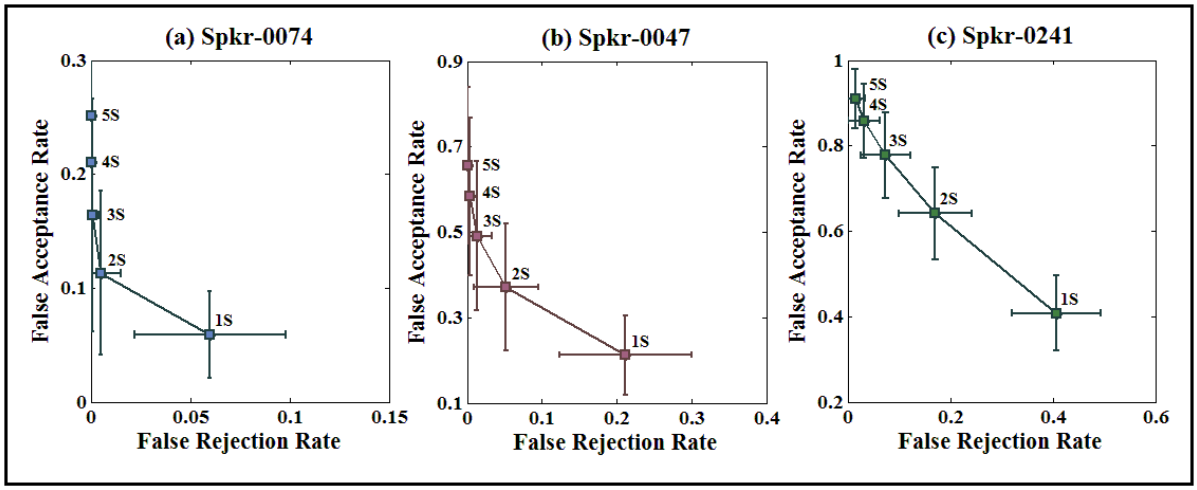

Figure 4.10 Multi-Instance Fusion Error Rates for three speakers from *SET-1* (a) Spkr-0074, (b) Spkr-0047, (c) Spkr-0241

The multi-sample fusion evaluated here is complementary to multi-instance fusion explained in previous subsection. As discussed earlier, the fusion performance depends on the base classifier errors that are considered speaker dependent. With the fusion of multiple samples, the number of *false acceptances* increases whereas the *false rejections* decrease irrespective of the base classifier performances. For example, fig. 4.10 represents the error rates for the sequential *'OR'* fusion of five samples for three speakers with good (Spkr-0074), average (Spkr-0047) and worse (Spkr-0241) base performances in *SET-1*. Similar results for rest of the speakers from *SET-1* are presented in the fig. 4.30. For each speaker, the points on each curve represent the mean error rates (FRR & FAR) with standard deviation for sample combinations. The increase in FAR is directly and FRR is indirectly proportional to the number of samples used for fusion. The fusion of two samples results in reducing the FRR by 16% (23.6%) and increasing the FAR by 15.9% (23.3%) for the Spkr-0047 (Spkr-0241). However, the increase in FAR is greater than decrease in FAR when the number of samples used for fusion are increased. For example, the fusion of five random samples results in 21.1% (44.4%) decrease in FRR and 39.1% (50%) increase in FAR for Spkr-0241 (Spkr-0047). The fusion performance for an application thus depends on base performance and the number of samples to be used for fusion.

As discussed in previous section, the factors such as data used for training and testing, dataset size and session variability affect the base classifier performance of a speaker. For disjoint test data, base performances are either similar or different (fig. 4.9). The fusion of multiple instances (digits) results in error rates with certain standard deviation proportional

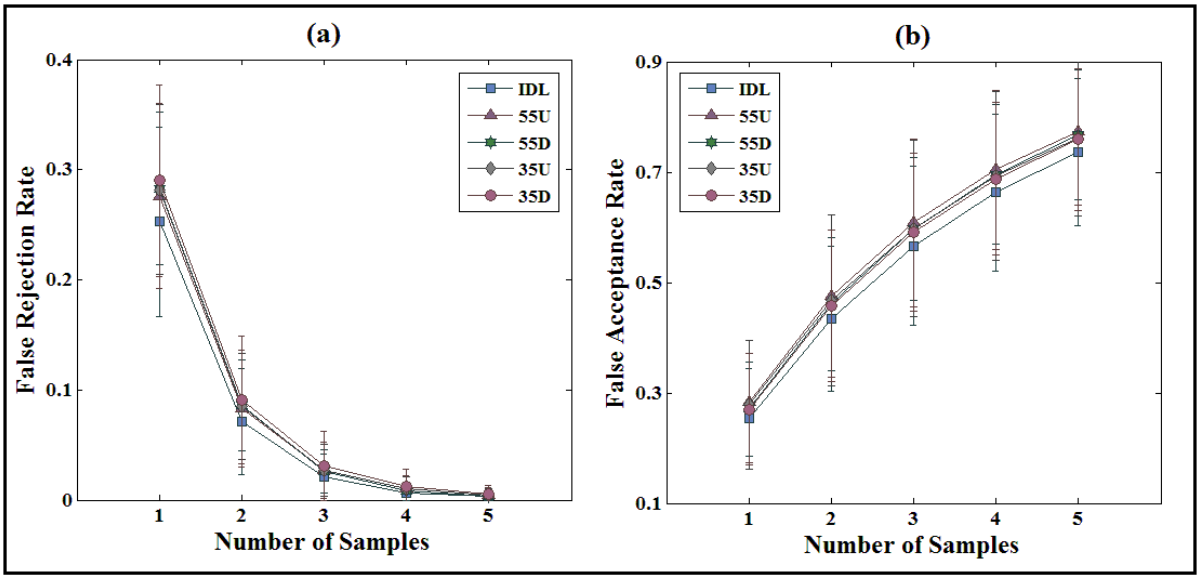

Figure 4.11 Multi-Instance Fusion Error Rates for four training sets of three speakers from *SET-1* (a) Spkr-0074, (b) Spkr-0047, (c) Spkr-0241

to base performances of a speaker across different datasets. The same results can be extended for overlapped test datasets (fig. 4.31). The use of different training models for verification results in different base classifier performances for the same dataset.

Figure 4.11 presents the error rates for multi-sample fusion of an instance from the same test-dataset with different training models (one model for each noise condition from *SET-3*). The multi-sample fusion performance is enhanced by selecting the training models with minimum error rates (for each digit or all digits) when the base performances are significantly different for each training model. However, the models with the comparably lower error rates for different training models do not always ensure minimum fusion error rates, as with the case with multi-instance fusion. The IDL model is shown to have better base performance compared to others and so the fusion error rates are lower for these models (fig. 4.11). The deciding factor for fusion performance, i.e., the complementary information from the decisions of different instances, is discussed in detail in next chapter. Though the analysis here is explained for multiple training models for the same dataset, the conclusion can be extended for the use of multiple training models on different datasets (fig. 4.32 presents the fusion error rates for different datasets verified on different training models for three speakers from *SET-1*).

The choice of threshold criteria used for verification can also significantly vary the base classifier performance and in turn the fusion performance. The mean error rates for the

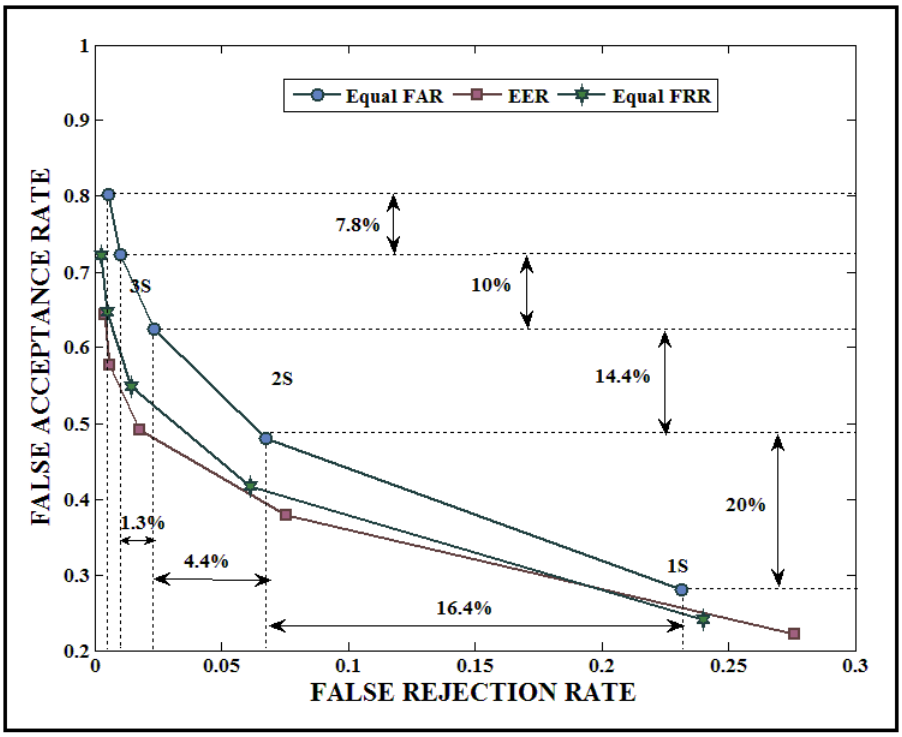

Figure 4.12 Error rates for multi-sample fusion of three different threshold criteria for Spkr-0047 from *SET-2*

Spkr-0047 from *SET-2* with three different threshold criteria (*EER* for each digit, *Equal FRR* and *Equal FAR* for all digits) are represented in fig. 4.9. The difference in fusion performance of a dataset for the variation in decision thresholds is observed to be minimal when the mean error rates for base classifiers are similar. With multi-sample fusion, the decrease in FRR reduces with an increase in the number of repeated samples, i.e., in fig.4.12 the FRR (threshold selected using *Equal FAR*) decreases by 16.4% when two samples are combined whereas this decrease in FRR reduces to 4.4%, 1.3% and 0.4% for three, four and five sample combinations respectively. Similarly, the increase in FAR decreases with an increase in number of samples fused. For example, the increase in FAR is 20%, 14.4%, 10% and 7.8% for two, three, four and five samples.

Although the results presented in this subsection are for random repetition of samples, the fusion of adaptive samples also decreases the FRR and increases FAR. The next section presents the difference in error rates for the fusion of random samples and adaptive samples.

4.4.2 Adaptive Samples

In real-life scenario, the number of samples required will be far less for true speaker acceptance, as the speaker will be good at adapting to his/her model. An impostor, on the other hand, may not be very good at adapting to claimant model and thus requires more samples for verification. For the evaluation of multi-sample fusion using adaptive samples, the speaker is allowed to mimic a word or phrase and try to sound as much like the claimant speaker as possible. This type of data is limited in the existing databases, for example, the CSLU database has one utterance where the speaker is asked to mimic a single sentence. The other approach that can be used for evaluation of adaptive sample fusion is the use of voice conversion/speech transformation tools. The Voice Conversion Matlab® Toolkit [180] is used to obtain adaptive samples for client and impostor verification. This method is described in the previous chapter (section 3.5.2.2).

For client-to-client conversion, the parameters are trained on the data (source and target) from the same speaker. For impostor-to-client conversion, the estimated parameters are trained on the source data from an impostor and target data from a client. The parameter estimation is performed for each digit of a speaker independently. For client-to-client conversion, the parameters are trained on the data (source and target) from the same speaker. For impostor-to-client conversion, the estimated parameters are trained on the source data from an impostor and target data from a client. For each sample that is rejected, a transformed sample is generated. It is also possible that the conversion technique will be good in transforming certain digits more accurately than other digit models, thus resulting in different error rates for isolated digits. It is to be noted here that certain samples accepted initially can be rejected after conversion. The performance of the system for fusion of adaptive samples is tested by progressively increasing the number of samples used for verification. If a speaker is accepted by 'i^{th} sample' then the subsequent samples (i+1, i+2 ... m) need not be used for verification and the fusion performance is thus independent of decisions from these subsequent samples.

Figure 4.13 presents the DET curves for adaptive samples (1st, 2nd and 3rd samples) of isolated digits of speakers (Spkr-0074, Spkr-0047 and Spkr-0241) from *SET-1*. The speaker verification performance for the adaptive samples is shown to be significantly different across the digits. The adaptability of the digit sample here depends on the training data used

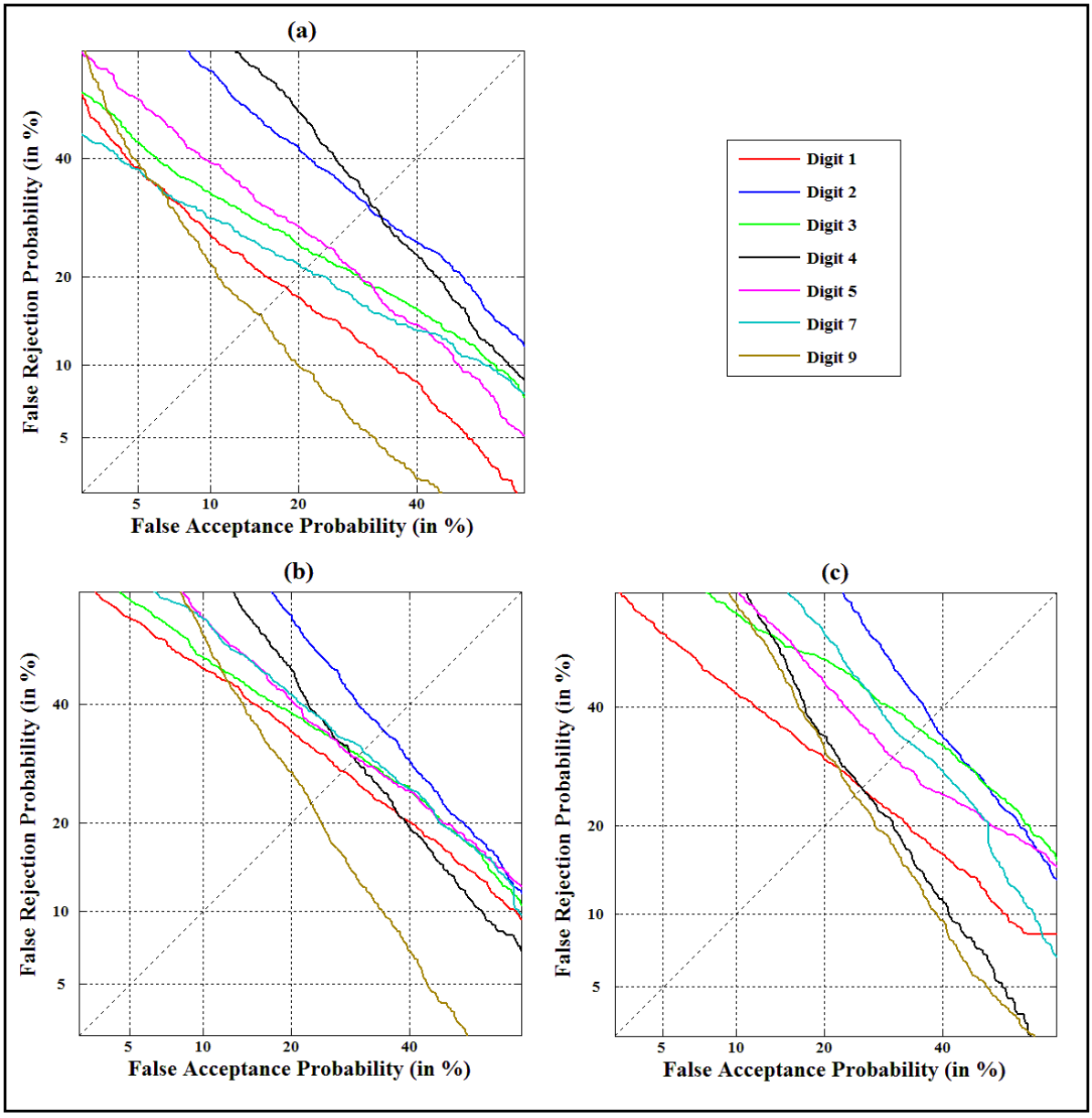

Figure 4.13 DET plots for verification of isolated digits for (a) Single Sample, (b) Two Adaptive Samples and (c) Three Adaptive Samples

for the parameter estimation of that particular digit. As a result, the client-to-client converted digit samples that are highly adaptive results in lower number of false rejections compared to the source speech samples. Similarly, highly adaptive impostor-to-client converted samples results in higher number of false acceptances. The Equal Error Rate (EER) for the isolated digits thus depends on the conversion parameters. For example, the use of converted adaptive samples (2nd and 3rd samples) decreases the EER for digit models *'Four'* but the EER increases for digit model *'Nine'*. The EER for repeated adaptive samples is lower when the decrease in FRR is higher than the increase in FRR and vice versa. For clarity of fusion

analysis, the repeated samples are obtained for even the accepted samples of a digit, i.e., the conversion technique is employed for both accepted and rejected samples by the base classifier. As the conversion process mainly depends on the PSOLA technique, the accepted samples when converted may also be rejected by the classifier.

Figure 4.14(a) represents the DET curves for the pooled verification tests of adaptive samples. The DET plot for verification of original samples is same for both random and adaptive samples. For random samples, the subsequent samples are randomly selected from the remaining dataset and so the DET curves are the same for any number of random samples. On the other hand, the DET curves for subsequent adaptive samples are different. When adaptive samples are individually tested, the false rejects decreases and false accepts increases compared to the errors for original samples. The Equal Error Rates (EER) for the second and third samples when compared to first sample is not low (fig. 4.14(a)).

The impact of multi-sample fusion on overall performance is better explained using the DET curves for individual speakers. Figure 4.14(b) shows the curves for the tests performed on samples from three speakers with good (Spkr-0074), average (Spkr-0047) and worse (Spkr-0241) performance. For Spkr-0241, the EERs for second and third samples are lower than first sample tests. This is because the decrease in FRR is higher than the increase

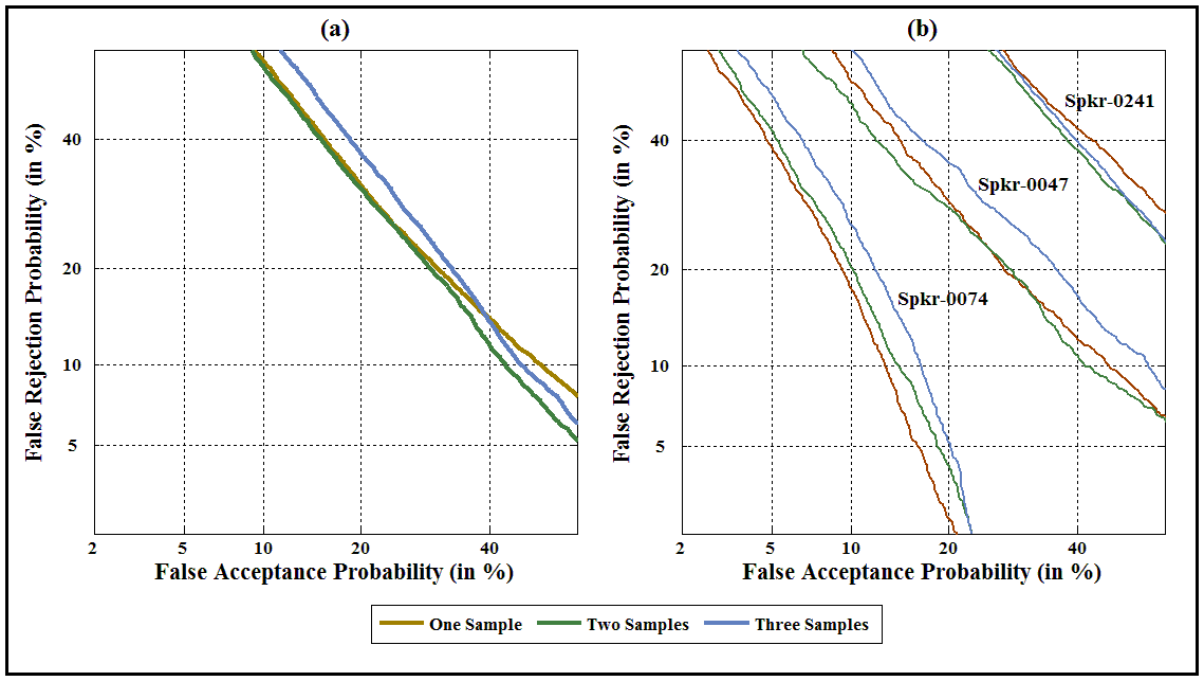

Figure 4.14 DET Curves for the speaker verification performance of tests performed on (a) all test speakers (pooled results) and (b) individual test speakers

in FAR when multiple samples are fused. Whereas, for Spkr-0074, the increase in FAR is significantly higher than the decrease in FRR (Here, the false rejects for the 2nd adaptive sample fast reaches to zero because of limited number of tests whereas the false accepts increases with each successive sample). Therefore, the overall performance of the system improves when the decrease in FRR for multiple samples is higher than increase in FAR.

Irrespective of the verification error rates for individual adaptive samples, the fusion of these samples is observed to have same effects as that of fusion of random samples. Table 4.2 presents the mean error rates with standard deviation for the fusion of two and three adaptive samples. The fusion of multiple samples, for both random and adaptive samples, is shown to increase FAR and decrease FRR. The false rejects are lower when a client speaker tries to adapt with each additional sample for verification. When an impostor tries to use an adaptation technique, the false accepts is higher than random repetition of samples. The total error rates for adaptive samples are higher than for repeated random samples, as the reduction in FRR for adaptive samples is lower than reduction in FAR.

The results for adaptive samples here are calculated under the assumption that an impostor has knowledge of the target speech data and the underlying information regarding the speaker verification system, i.e., the parameterisation and modelling parameters and techniques. As obtaining this information in general is difficult, the increase in number of false acceptances for multi-sample fusion can be less than the results presented in this section for adaptive samples. The analysis in this section holds good for actual adaptation by a client and/or imitation by an impostor instead of conversion tools. The true speaker utterance in this case has a better chance in being accepted rather than an imitation by an impostor thereby improving the chances of reducing the total error rate for the multi-sample fusion.

Table 4.2 Verification Error Rates for fusion of random and adaptive repetitive samples

	Random Samples		Adaptive Samples	
	FRR	FAR	FRR	FAR
1-sample	$0.226^{\pm 0.16}$	$0.226^{\pm 0.16}$	$0.226^{\pm 0.16}$	$0.226^{\pm 0.16}$
2-samples	$0.076^{\pm 0.06}$	$0.389^{\pm 0.24}$	$0.061^{\pm 0.05}$	$0.393^{\pm 0.23}$
3-samples	$0.029^{\pm 0.04}$	$0.485^{\pm 0.28}$	$0.019^{\pm 0.01}$	$0.492^{\pm 0.27}$

The use of multiple samples, either random or adaptive, results in reducing the FRR at the cost of increase in FAR. This effect of multi-sample fusion is complementary to that of multi-instance fusion (section 4.5.2) where the FAR reduces with increase in FRR. This increase in error rates for the fusion schemes (FRR for multi-instance and FAR for multi-sample fusion) is reduced by combining the multi-instance and multi-sample fusion schemes. This scheme is analyzed empirically in the next subsection.

4.5 Fusion of Multi-instance and Multi-sample schemes

OVERVIEW: The proposed architecture is based on the sequential integration of multi-instance and multi-sample fusion schemes. This method is analytically shown (section 2.7.3) to improve the performance and allow a controlled trade-off between false alarms and false rejects when the classifier decisions are statistically independent. When the proposed architecture is applied for text dependent speaker verification using Hidden Markov Model (HMM) based digit dependent speaker models, the improvement in fusion performance is achieved irrespective of the number of instances and samples used for fusion. The fusion performance is then evaluated for with tuning of parameters, n classifiers/instances and m attempts/sample.

The proposed architecture is based on the integration of multi-instance and multi-sample fusion schemes. This architecture (Figure 4.15) is used for testing (test dataset) in which the maximum permissible number of repeated samples '*m*' and the number of instances '*n*' are fixed prior based on the error rates obtained from the development dataset. In this system, the speaker presents an input test utterance x_{ij} *(i=1, 2 ...n, j =1, 2 ...m)* and the classifier ζ_i (here HMM) makes a decision to either accept or reject the claimed identity. The final decision of the proposed system is to accept the claim only if the speaker is accepted by '*n*' classifiers in sequence within the maximum number of allowed multiple samples '*m*'. The claim is rejected if the speaker is not able to get accepted at any one of the classifier within the allowable number of multiple attempts '*m*'.

The protocol for the performance evaluation of this fusion scheme is described in section 3.5.2. An ***instance***, here, refers to a ***digit*** that is modelled using a HMM to represent a speaker's characteristics. Therefore, *multi-instance fusion* is referred as the *combination of digits*. A ***sample*** for an *instance*, here, refers to an ***utterance*** of a *digit* that is used for speaker

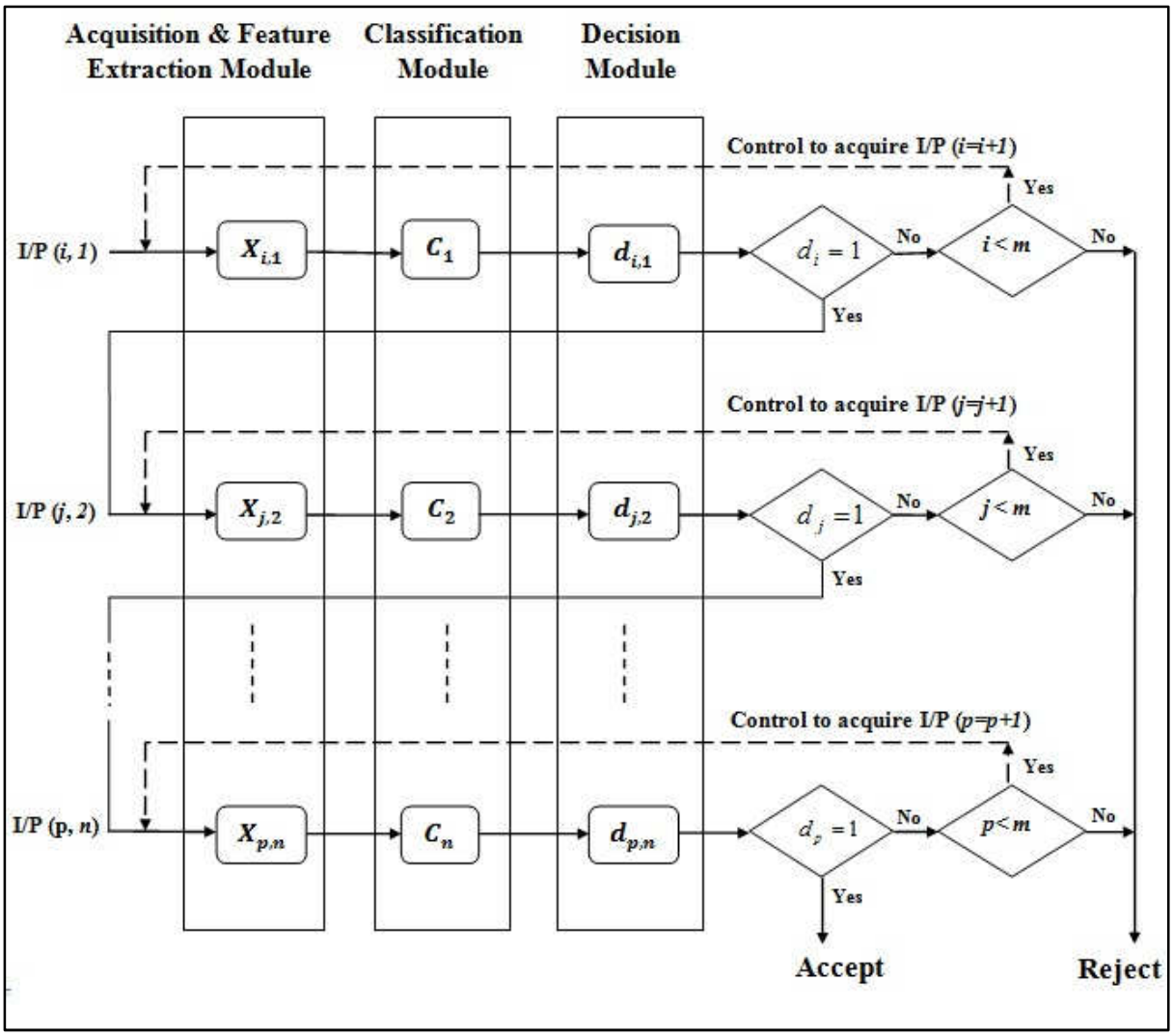

Figure 4.15 The architecture of a multi-instance and multi-sample fusion scheme with *'OR fusion'* of *'m'* repetitive samples and *'AND fusion'* of *'n'* classifiers

verification by a HMM that represents the characteristics of a speaker. Therefore in this section, *multi-sample fusion* is referred as the *combination of decisions from multiple utterances of a digit*. The scheme is evaluated for speech data from *SET-1*, *SET-2* and *SET-3*. In this work, the analysis is presented using the pooled results for all the test datasets.

The fusion performance of the system is tested by progressively increasing the number of instances/digits used for verification where each instance is allowed with multiple samples. The false rejection and false acceptance rates of the proposed fusion scheme for test datasets of *SET-1*, *SET-2* and *SET-3* are presented in fig. 4.16. The point (*n*, *m*) = (1, 1), in fig. 4.16, represents the mean error rates for tests on isolated digits (*instance*, *sample* = 1, 1). The other points also represent error rates for parameters (*instance*, *sample*) combinations. The FAR decreases with the increase in digits/instances used for fusion and FRR decreases

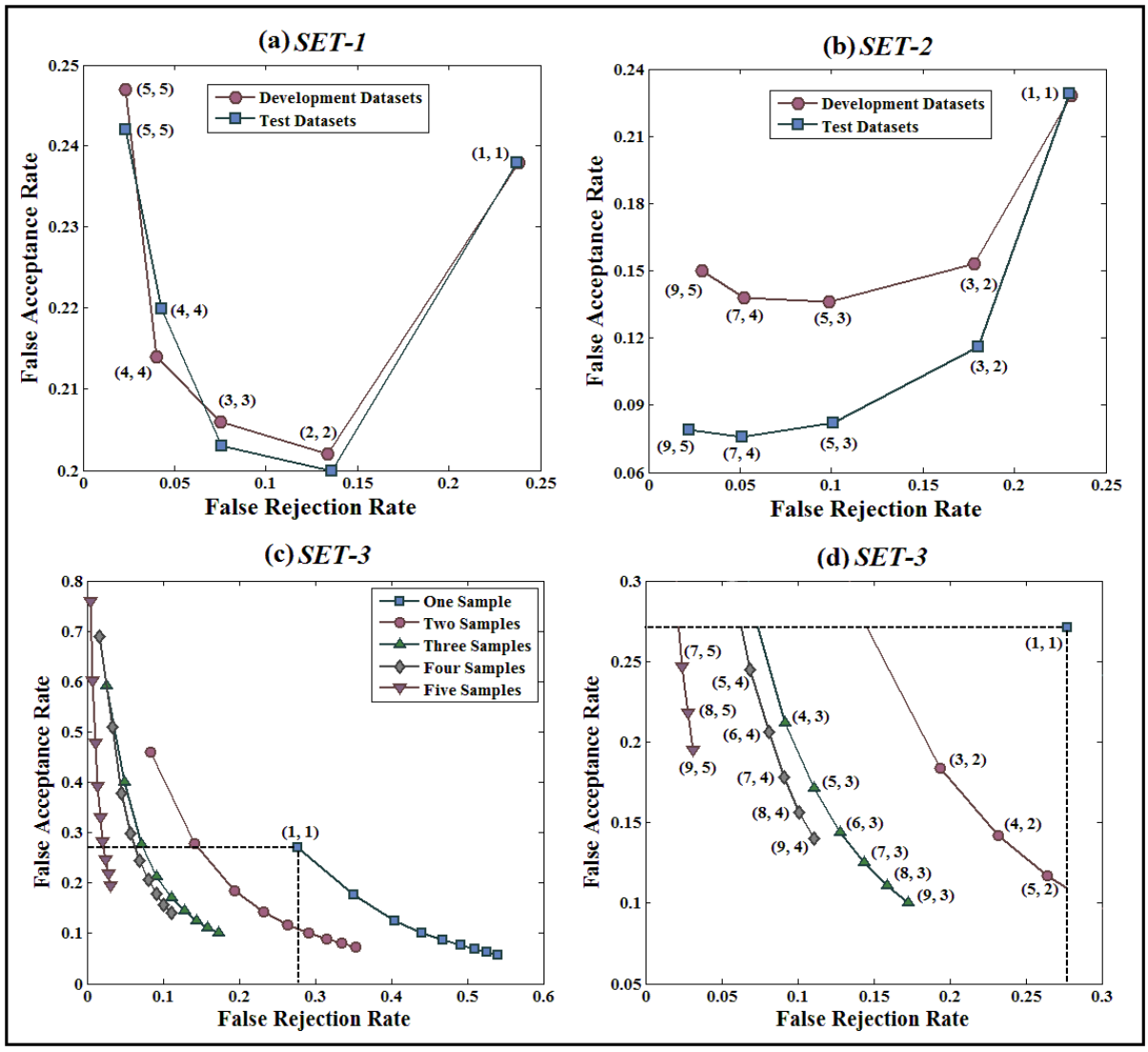

Figure 4.16 Verification error rates for the proposed multi-instance and multi-sample fusion of test datasets from (a) *SET-1* (n=m), (b) *SET-2* (n>m), (c) *SET-3* ($\forall n, \forall m$) and (d) *SET-3*(errors lower than *(1, 1)*)

when multiple samples are allowed for each of these digits. In fig. 4.16 (a), the FRR and FAR errors for progressive increase in digits and samples are represented for tests of *SET-1*.

The reduction in errors, FRR and FAR, is monotonic initially but soon reaches saturation. In fig.4.16(a), the performance increases until the combination of five instances and five samples for which fusion FAR is higher than base FAR of isolated digits. The fusion errors for (5, 5) here is catastrophic as the fusion TER is higher than base TER as the increase in fusion FAR here is greater than decrease in fusion FRR. Better error rates are obtained by ensuring that the digits used for fusion are more than the samples.

In fig. 4.16(b), the error rates are represented for (*n*, *m*) combinations where $n>m$ and $n>1$, $m>1$ for tests on *SET-2*. The total error rate for these selected combinations is observed to decrease monotonically. The decrease in FAR or FRR for these combinations may not be monotonic though the errors are less than the base error rates for isolated digits. The best parameter combination is determined based on the application requirement, i.e., user convenience or high security. In fig. 4.16(c), the error rates for all possible parameter combinations with $\forall n$ and $\forall m$ are shown for *SET-3*. Each curve (in fig. 4.16(c)) represents the false accepts and false rejects for use of multiple samples and the points on each curve represents the digit combinations with progressively adding instances from top left to the bottom right of figure. The points below the line for the data point (*1, 1*), shown clearly in fig. 4.16(d), have improved fusion performance (e.g., the points (*3, 2*), (*4, 2*), (*4, 3*)...) has FRR and FAR lower than for point *(1, 1)*).

By tuning the parameters (*n, m*) to any value that falls below the area of lines for the data point (*1, 1*), both the verification error rates can be arbitrarily reduced with a trade-off in verification time. For example, the fusion of three digits and three samples (3, 3) for each digit of *SET-1* results in reducing the false rejection and false acceptance rates by 7.6% and 20.3% respectively. When nine digits and three samples *(9, 3)* are combined, the FRR reduces from 23% to 2.2% and FAR decreases from 22.9% to 7.9% for *SET-2*. For the parameter combinations below the line for the point (1, 1), the lowest FRR of ***2.4%*** can be achieved for (7, 5) where the FAR is reduced to 24.7% for *SET-3*. Similarly, the lowest FAR of ***0.1%*** can be obtained for (9, 3) that results in FRR of 11.8%. The improvement in fusion performance for the same parameter combination (*n, m*) is observed to be different across the *SETs*. This difference is mainly because of the variations in base performances for isolated digits of individual *SETs*.

From sections 4.3 and 4.4, the performance improvement of multi-instance and multi-sample fusion schemes is shown to be dependent greatly on base digit performances and therefore the performance of proposed architecture depends on base performances. Figure 4.17 presents the mean error rates for four combinations (instance, sample) of three speakers from *SET-1*. The fusion of two digits with two samples (2, 2), three digits with three samples (3, 3) and four digits with four samples (4, 4) are compared to base performances (1, 1). Similar curves for rest of the speakers from *SET-1* are presented in the fig. 4.33. The verification errors, i.e., false accepts and false reject, are arbitrarily reduced (mostly for all speakers) with trade-off in verification time. However, combinations with the increase in

digits/samples, the fusion performance can be worse than base classifier performances that lead to catastrophic fusion. The fusion of multiple instances/samples for speakers with relatively worse base classifier performance (for example in fig 4.33 the 2D-2S for speakers 0124 & 0241) results in higher FAR/FRR than the base classifier false acceptances (1, 1). The improvement in fusion performance is achieved, irrespective of base performances, by ensuring that the number of instances (n) are higher than the samples (m) used for fusion ($n > m$). (evident from fig. 4.16(b) and also in fig. 4.17 (c) where the error rates for digit combinations ($n > 4$) with three samples shows an improvement in performance compared to the base classifiers for the Spkr-0241).

In fig. 4.17(a), the error rates for Spkr-0074 reaches zero (i.e., false rejects equal to zero when samples ($m > 3$) or false accepts equal to zero when ($n > 4$)) because the number of tests performed are not sufficient to present the error rates at required precision. The tuning of parameters to (4, 3) i.e., the combination of four digits and three samples, reduces the FRR and FAR for Spkr-0074 to (≈) 0.08% and 0.03% respectively. The error rates for Spkr-0241 (fig. 4.17(c)) are lowered to 39.6% (FRR) and 28.4% (FAR), when three digits are combined with two samples for each digit. This improvement here is due to 6.5% and 21.1% reduction in FRR and FAR for Spkr-0241. The fusion of five digits with four repeated samples at each digit reduces the error rates by 10.5% (FRR) and 10.3% (FAR). It is thus demonstrated that there is potential to improve the performance of even weaker classifiers by combining them in this manner. The verification error trade-off here depends on tuning the parameters of instance and samples (n, m) used for verification.

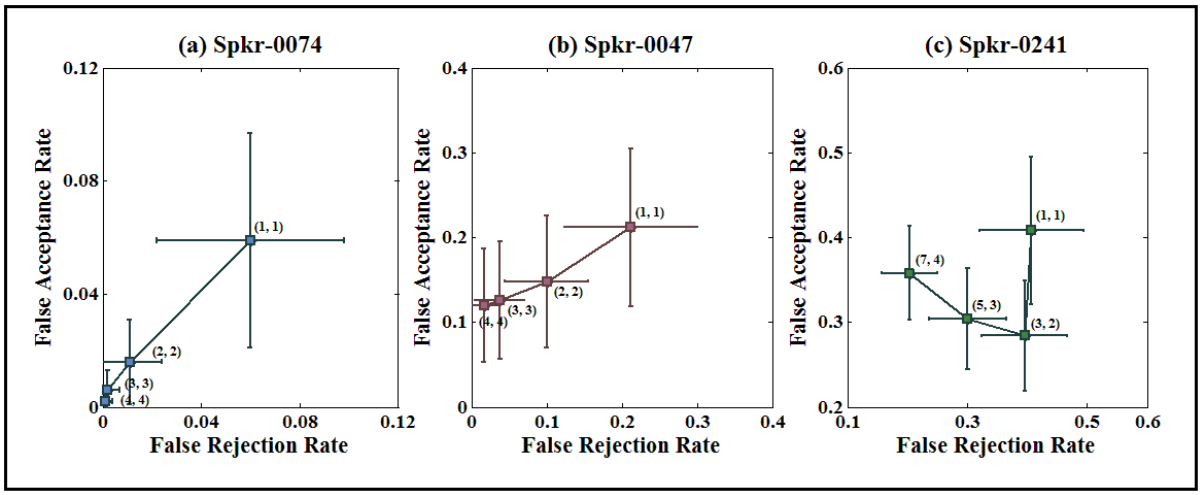

Figure 4.17 Proposed Fusion Error Rates for (a) Spkr-0074, (b) Spkr-0047, (c) Spkr-0241 from *SET-1*

The fusion of multiple samples, for both random and adaptive samples (section 4.4), increases FAR and decreases FRR. The false rejects are lower when a client tries to adapt with each additional sample for verification. When an impostor tries to use an adaptation technique, the false accepts (FAR) is higher than random repetition of a sample. This increase in FAR can be reduced for the combination of digits. The total error rate (TER) for the combination of digits with adaptive samples is reduced as, in general, the decrease in false rejects is higher than increase in false accepts for real scenarios.

The verification error rates for the fusion of digits with random and adaptive samples are presented in fig. 4.18(a) for pooled results of three speakers from *SET-1*. The FRR for adaptive samples is observed to be lower than random samples whereas FAR is lower for random samples. The points on curves for random and adaptive samples below the line for point (1, 1) represent the error rates that are lower than base error rates (EER of 22.6%). For example, the point (7, 3) represents the FRR (FAR) of 9.5% (9.8%) and 15.7% (7%) for adaptive and random samples respectively. Here, the FRR are lower for adaptive samples whereas the FAR are smaller for random samples. The parameter combinations with low FRR and FAR compared to base performance (1, 1) can be different for random and adaptive samples. Whereas certain parameter combinations (such as (3, 2) and (4, 3)) can have improved fusion performance - arbitrary reduction in FRR and FAR - for both random and adaptive samples. The overall fusion performance, represented using total error rate (TER) in fig. 4.18 (b)), is observed to be better for adaptive rather than random samples where $n>3$.

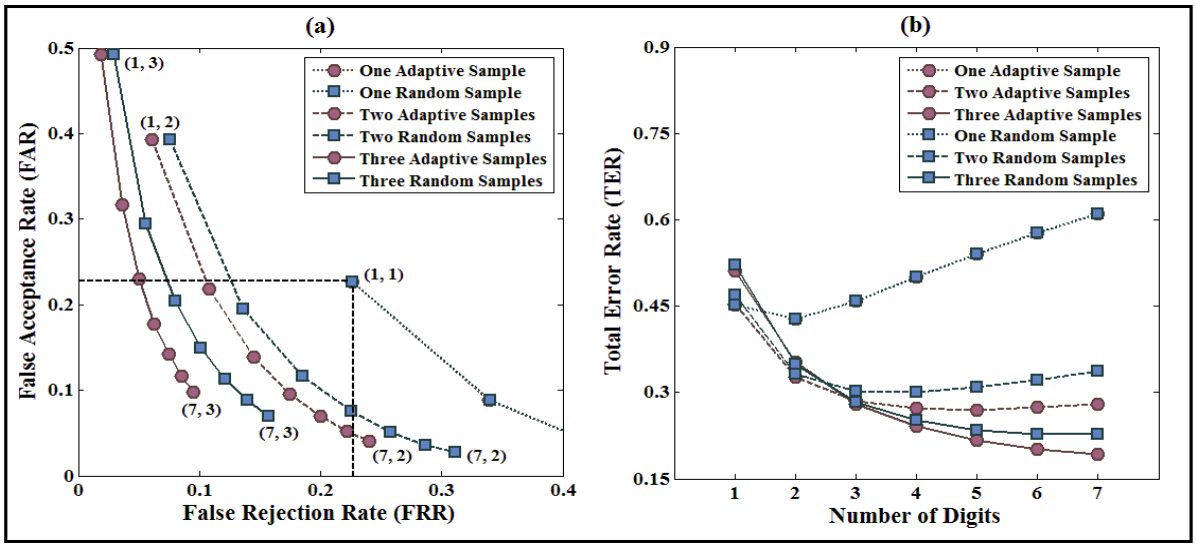

Figure 4.18 Sequential Decision Fusion of Adaptive and Random Samples (a) False Rejection Rate vs. False Acceptance Rate (b) Total Error Rates

The total error rate (TER) at the point (7, 3) is lowest for adaptive samples rather than random samples (fig. 4.18(b)). Therefore, in real scenarios where client and impostor adaptive samples are used for verification, the proposed fusion scheme is employed for performance improvement with arbitrary reduction in false rejects and false accepts.

In addition to the nature of the repetitive sample, the choice of threshold selection criteria also affects the improvement in fusion performance. Figure 4.19 represents the error rated for proposed fusion of tune datasets with three threshold selection criteria EER (fig. 4.19(a)), Equal FAR (fig. 4.19(b)) and Equal FRR (fig. 4.19(c)) for Spkr-0047 (*SET-2*). The false acceptances and false rejections are reduced simultaneously, irrespective of the threshold selection criteria, by tuning the parameters - number of instances and number of samples. Nevertheless, the selection of Equal FRR as threshold criteria results in lower number of false acceptances and false rejections compared to the other cases (Equal FAR and EER). When Equal FRR criterion is used for threshold selection, the fusion parameters (*9*, *5*) reduces the FRR and FAR errors by 25.6% & 17.7% for (*9*, *5*). For the same parameter combination, the FRR and FAR errors are reduced by 19.5% and 6.2% for Equal FAR criteria and are lower compared to errors for criteria of EER (22.2% and 10.9%). Lower TERs are obtained for the fusion of digits where thresholds are selected using Equal FRR criteria as the decrease in false accepts for these combinations is higher than increase in false rejects. Further, for Equal FRR criteria the number of suitable combinations with better TERs is more compared to the threshold selection criteria with Equal FAR and EER.

In general, the fusion performance can be improved when multiple thresholds are considered for speaker verification. The sequential probability ratio test (SPRT) developed by Wald has been proposed in the literature that employs an upper and lower threshold to better determine the identity claim. The performance of the proposed method is compared for single and two verification thresholds (SPRT technique). The datasets used for this comparison considers the Equal FRR threshold estimation criteria. The two thresholds are selected for this dataset such that the base classifiers have equal FRR and equal FARs. For SPRT with high upper threshold, the number of false accepts decreases and the false accepts increases compared to the proposed fusion with single threshold. The increase in FRR can be lowered by allowing multiple samples at the cost of increase in FAR.

The error rates for the proposed fusion with single and two thresholds are presented in table 4.3. The increase in FAR for combination of multiple samples may outperform the affect of combining multiple instances that results in reducing the FAR (FAR value for (6, 5)

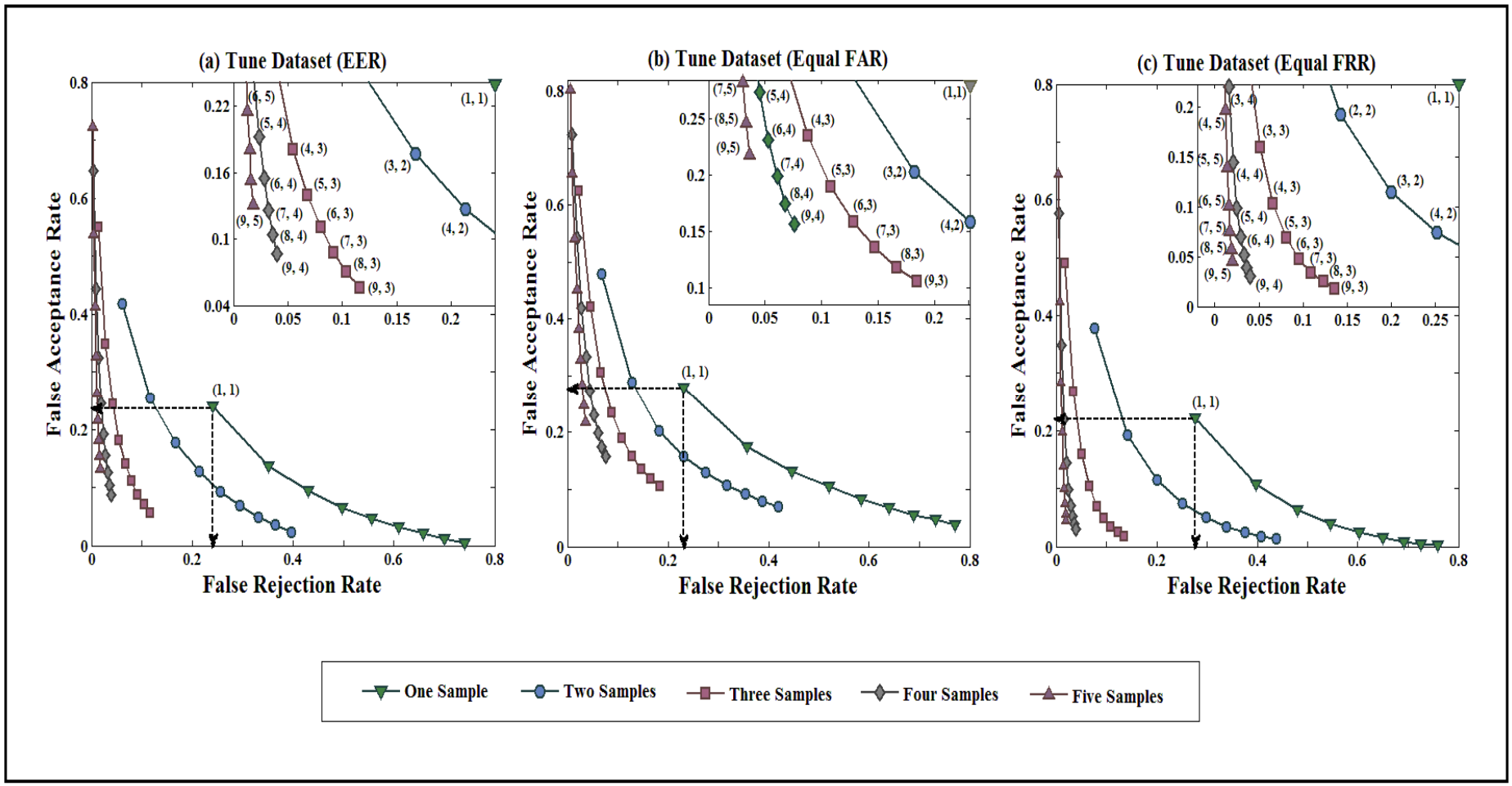

Figure 4.19 Verification error rates for proposed fusion of a tune dataset with different thresholds for speaker-0047 (*SET-2*)

Table 4.3 Error Rates for SPRT and Proposed Fusion Methods for speaker-0047 (*SET-2*)

	Sequential Probability Ratio Test (Two Thresholds)		Sequential Decision Fusion (Single Threshold)	
	FRR	FAR	FRR	FAR
1D-1S	0.276	0.020	0.276	$0.222^{\pm 0.126}$
3D-2S	$0.686^{\pm 0.133}$	$0.131^{\pm 0.005}$	$0.200^{\pm 0.012}$	$0.115^{\pm 0.069}$
4D-3S	$0.667^{\pm 0.132}$	$0.222^{\pm 0.005}$	$0.066^{\pm 0.009}$	$0.102^{\pm 0.060}$
5D-4S	$0.631^{\pm 0.099}$	$0.009^{\pm 0.005}$	$0.025^{\pm 0.003}$	$0.099^{\pm 0.055}$
6D-5S	$0.614^{\pm 0.071}$	$0.010^{\pm 0.004}$	$0.016^{\pm 0.003}$	$0.101^{\pm 0.049}$

is higher than for (5, 4) in table 4.3). Although the tuning of parameters *(n, m)* for SPRT method reduces both error rates, better trade-off between FRR and FAR is obtained with limited speech (less number of instances and samples) when single threshold is used verification.

The results demonstrate that it is possible to design a fused system with lower errors of both types compared to a single verification instance using a single sample. It is also possible to obtain controlled verification errors with trade-off in the time for computations required to perform multiple matches and make decisions with every sample and instance in the architecture. Assuming the computation time for an instance verification to be *t/n* seconds, the trade-off on using '*m*' multiple samples for '*n*' instances becomes the increase in total time for verification to an upper limit of '*mt*' [23]. However, the total verification time is often less than the upper limit. This is because, in general, the number of samples required by a true speaker to be verified correctly is far less than that of an impostor. Therefore, there is a possibility for the true speaker to be accepted before reaching the maximum number of attempts and so the verification time at each instance is mostly less than '*mt*'. Further, in a sequential system, if the classifier decides to reject a speaker at any of the intermediate stage, the processing of samples for the subsequent instances does not take place. So in the case of a reasonably performing classifier, the total verification time for '*p*' number of instances with '*m*' attempts is less than '*mt*' (i.e., $p*m*t/n < mt$, $p<n$). Hence, it can be considered that the false acceptance rate can be reduced arbitrarily without trading off the false rejection rate, at the expense of some increased time for a verification process.

4.6 Sequential Fusion of Multiple Information Sources

OVERVIEW: The proposed multi-instance and multi-sample biometric fusion scheme has been shown to improve performance with better trade-off between false accepts and false rejects in the previous section. Some recent studies on multibiometrics have shown that not all biometrics of the same modality perform equally well. Further, combining different biometric sources induces some drawbacks such as the increase in complexity of the system leading to a higher cost, longer verification time and lower user convenience. As the system cost increases with the changes in existing technology of the speaker verification system, the proposed scheme is tested for the integration of with other sources of information i.e., multiple models other than multiple instances or multiple samples. The results for the evaluation of such architecture are demonstrated in this section to determine the best nature of biometrics that can be used for fusion performance improvement under the assumption of independence between the biometrics.

The proposed system's requirement for additional samples is considered an inconvenience for clients in certain applications and so the number of samples allowed is restricted to minimum. Therefore, alternative methods are determined to obtain multiples decisions for an instance without changing the actual technological specifications for speaker verification. The multibiometric systems with multiple modalities or multiple algorithms require changes in technical aspects such as feature extraction and/or processing techniques. The other source for information can be from multiple sensors for acquisition of a sample. The existing databases have limited data for a speaker from multiple sensors. Therefore, the effect of other sources of information on proposed fusion is tested for multiple models, where each model can be created using different speech data for verification.

The architecture (fig. 4.8) for fusion of multiple samples can be extended to a fusion scheme where multiple models are used for reduction in false rejects instead of multiple samples. The fusion scheme is evaluated by selecting the subsequent model for the rejected sample randomly. The models trained for different noise conditions (IDL, 55U, 55D, 35U and 35D) in *SET-3* are used for evaluation of multi-model fusion with either multi-instance or multi-sample fusion schemes. This architecture is similar to the proposed sequential *'AND'* and *'OR'* decision fusion shown in fig. 4.15.

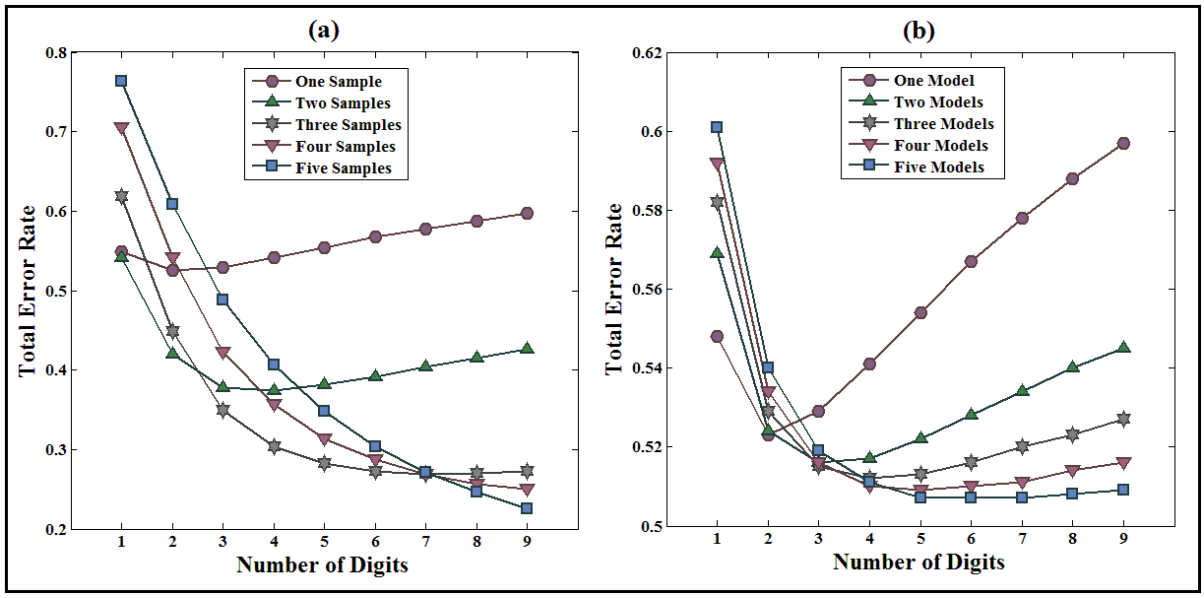

Figure 4.20 Total Error Rates for fusion of multiple instances with (a) multiple samples and (b) multiple models

For a speaker to be declared genuine at an instance (or spoken text) stage, it is considered sufficient if the utterance/sample for the digit is accepted by a single model, here IDL model. If the speaker is rejected, then the digit sample is verified for another randomly selected model. Acceptance decisions are logical *'OR'* and rejection decisions are logical *'AND'* for multiple models. Conversely, it is considered necessary in the sequential decision framework that a speaker be accepted by all instances in the sequence of decision stages. Acceptance is thus logical *'AND'* for multiple instances. If the speaker is rejected by any decision stage, the sequence terminates and thus rejection decisions are logical.

The total error rate for fusion of instances and samples are shown in fig. 4.20(a) whereas fig. 4.20(b) presents the error rates for fusion of multiple instances and multiple models for speakers from *SET-3*. The error rates are observed to be higher for multiple models compared to multiple samples. The FRR and FAR error rates for isolated digits (1, 1) are 27.7% and 27.1% respectively. The fusion of nine digits and five samples results in reducing FRR and FAR to 3.1% and 19.5% whereas the fusion of nine digits with five models reduces the errors to 36.2% and 14.7% respectively. The total error rates for fusion with multiple samples are mostly below 40% for instances greater than three whereas TER is greater than 50% for fusion of instances (>3) with multiple models.

When the above fusion scheme is modified such that multiple models are combined using sequential *'AND'* fusion with multiple instances/digits combined with sequential *'OR'*

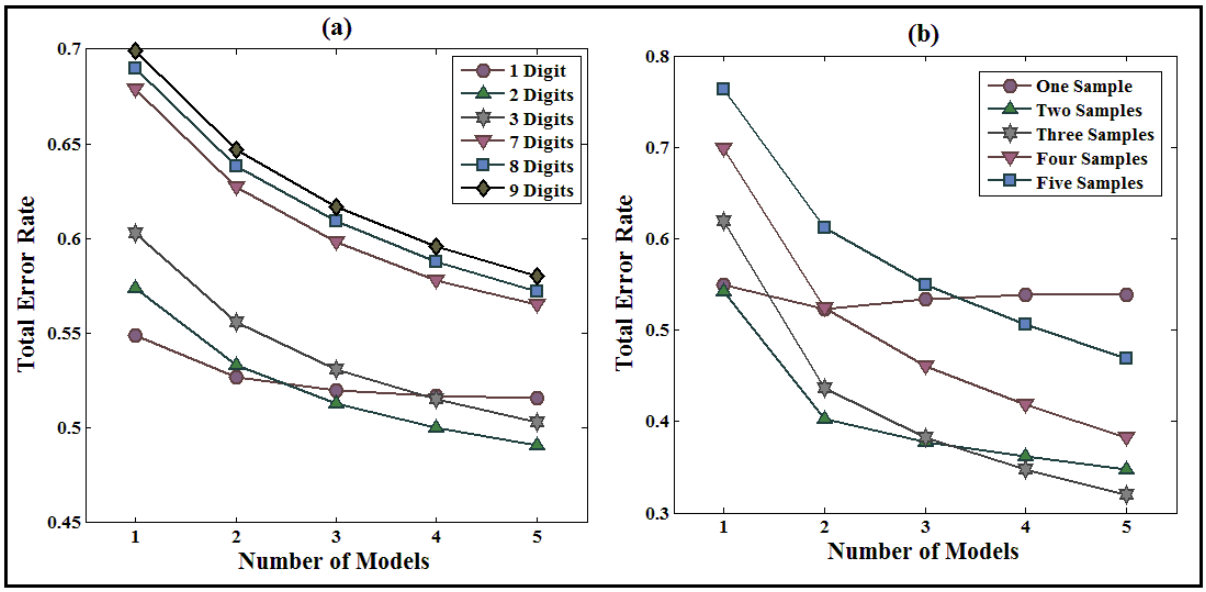

Figure 4.21 Total Error Rates for multi-model and multi-instance fusion scheme

fusion, the fusion results obtained are mostly considered be catastrophic (fig. 4.21). This scheme can be employed in scenario where each model has certain number of best digit combinations. If the speaker is rejected for a digit, then the next best for the particular model is verified. If the digit is accepted for the first model, then the speaker is verified for the next model and its corresponding best digit. The best digits for a model here are selected based on minimum total error rates. Figure 4.21(a) presents the TER for multi-model and multi-instance fusion scheme where the errors increase with the digits used for fusion. The TER for fusion is lower than base errors when only two and three best digit for are combined. But the error rates increases progressively with each addition of a digit for verification.

If the information from multiple instances, in the above scheme, is replaced with multiple samples then the fusion architecture is based on integration of multi-model and multi-sample fusion. The TERs for this fusion scheme are presented in the fig. 4.20 (b). Although the performance for fusion of multiple samples is observed to be better than for multiple instances when integrated with multi-model fusion, the proposed scheme - multi-instance and multi-sample fusion, achieves best fusion performance. This difference in performance is mainly due to the complementary information from biometric sources i.e., the error rates are reduced for the scheme that better exploits the complementary information between the biometric sources [191]. Therefore, the multiple instances and multiple samples are considered as the best information sources for the proposed decision fusion scheme.

4.7 Error Rates for Fixed Fusion Rules

OVERVIEW: The proposed fusion scheme that integrates the decisions from multiple instances and multiple samples has been shown to reduce the total error rates. Nevertheless, the design of the hybrid multibiometric system is shown to depend on the optimal classifiers and optimal combination method. This section thus compares the performance of proposed fusion with the existing decision-level combination methods. The methods used for evaluation comparison are Max, Min, Median, OR, AND and the Majority voting rules. The aim is to determine the optimum fusion method from amongst these techniques best suited for fusion of multiple instances and samples for Text-Dependent speaker verification. The total error rates for the proposed fusion scheme are observed to be lower compared to other fixed fusion schemes.

Decision fusion techniques are divided into fixed and trained rules. The choice between fixed and trainable rules is based on factors such as the size of the data set (large datasets are necessary for using trainable rules), size of the classifier combination pool (small ensembles should be preferred for trained rules), the degree of balance in classifier performances (fixed rules usually work well with balanced classifiers), etc. The fixed fusion rules can outperform the trained rules, such as weighted averaging and behaviour knowledge space, for a small validation set with classifiers of similar accuracy [192]. The performance of fixed fusion rules [193], such as *MAX*, *MIN*, *MEDIAN*, *AND*, *OR* and Majority Vote rules are statistically compared with proposed fusion.

The final decision of acceptance or rejection is the combination of individual *'n'* instances and *'m'* samples $d_{ij} \mid \forall_i (i = 1,2,3....n; j = 1,2,3,...m)$. The fusion technique used can be expressed using the generalised function as:

$$d_{COM} = f_{COM}\left(d_{11}, d_{22}, ..., d_{nm}\right)$$

The strategies used for combination of decisions here are based on fixed rules - *MAX*, *MIN*, *MEDIAN*, *AND*, *OR* and Majority Vote. These functions are defined as:

1. $d_{\min} = \min_{ij}\left(d_{ij}\right)$,

2. $d_{\max} = \max_{ij}\left(d_{ij}\right)$,

3. $d_{\text{med}} = \text{median}_{ij}\left(d_{ij}\right)$ and

$$d_{MajVote} = \begin{cases} 0, & for\ m \geq k\ decisions\ with\ d_z = 0 (z = 1, 2, .., m) \\ 1, & for\ m \geq k\ decisions\ with\ d_z = 1 (z = 1, 2, .., m) \end{cases}$$

4.

$$where\ \ k = \begin{cases} \dfrac{(i+j)}{2} + 1, & if\ \text{(i+j)}\ is\ even \\ \dfrac{(i+j+1)}{2}, & otherwise \end{cases}$$

The architecture of the proposed scheme considers the fusion of outputs from multiple instances and multiple samples using fixed rules i.e., *'AND'* and *'OR' Rules*. The decisions from multiple instances are combined using *'AND Rule'* where the claim is accepted only if the speaker is accepted for all instances. Whereas the multiple sample decisions are fused with *'OR Rule'* and thus the speaker claim is accepted, if accepted at any one sample for an

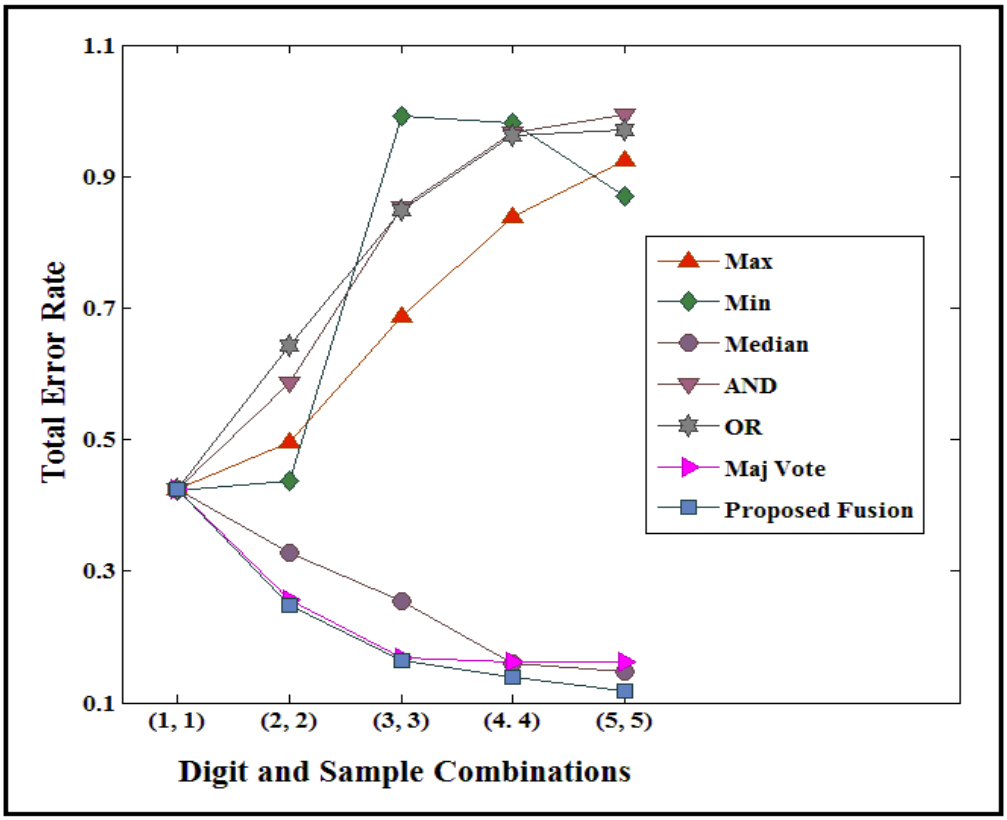

Figure 4.22 Total Error Rates for fixed fusion techniques of Spkr-0047 for *SET-1*

instance. Figure 4.22 presents the total error rates (TER) for different trained fusion techniques with comparison to the proposed fusion scheme. The TER for the fixed rules *AND*, *OR*, *Max* and *Min* Rules increase with number of samples and instances used for fusion. The errors for majority voting and median rules decrease with increase in number of instances and samples used for fusion. Nevertheless, the decrease in TER for majority vote reaches saturation and then increases with each addition of an instances and/samples. The total error rates for the proposed fusion scheme are observed to be lower compared to other fixed fusion schemes.

4.8 Comparison of Ideal and Experimental Error Rates

OVERVIEW: The proposed architecture, in section 2.7.3, based on the sequential integration of multi-instance and multi-sample fusion schemes is analytically shown to improve the performance and allow a controlled trade-off between false alarms and false rejects when the classifier decisions are statistically independent. The architecture when evaluated for the text-dependent speaker verification, in section 4.5, validated the improvement in fusion performance. As the assumption of statistical independence between decisions may not be always valid, this section validates if the theoretical/ideal error rates, calculated using the equations developed under this assumption, are equal to the experimentally obtained error rates. The comparison between the ideal and experimental false rejection rates and false acceptance rates are presented for multi-instance fusion, multi-sample fusion and integration of both these fusion schemes.

In real world applications, the verification system may set initial acceptable values for the number of false rejections and false acceptances. These error rates can be theoretically estimated using fusion parameters, i.e., the number of instances and samples, used for verification. These parameters are designed using the base error rates (FRR and FAR) of the isolated digit. The theoretical estimation of errors for multi-instance and multi-sample fusion schemes are explained in section 2.7.3. When false rejection rate 'ρ' and false acceptance rate 'α' are the same for '*n*' instances and '*m*' samples, the expressions for sequential fusion of independent decisions can be given as [23]:

$$\left.\begin{array}{l} \text{Multi-Instance Fusion}: \alpha(n,1) = \alpha^n;\ \ \rho(n,1) \approx n\rho\left(\text{when } \rho << 1\right) \\ \text{Multi-Sample Fusion}: \alpha(1,m) \approx m\alpha\left(\text{when } \alpha << 1\right);\ \ \rho(1,m) = \rho^m \\ \text{Multi-Instance \& Multi-Sample Fusion}: \alpha(n,m) \approx \left(m\alpha\right)^n;\ \ \rho(n,m) \approx n\rho^m \end{array}\right\} \quad (4.1)$$

The assumption of independence holds good for fusion of different biometric characteristics and may not be true for multibiometric systems with multiple sources from a single modality. In case of statistically independent decisions, the estimated ideal error rates (4.1) are equal to the experimental error rates. The dependence between the decisions can thus be determined by analysing the difference between these two errors.

A test for significance can compare the theoretically calculated error rates to the experimental error rates. A measure of how well the ideal and experimental error rates agree can be expressed in terms of a probability (p-value) [194]. The statement being tested is called the null hypothesis. The null hypothesis here assumes that there is no significant difference between the means of ideal and experimental error rates. The test of significance usually assesses the strength of evidence against the null hypothesis. When the p-value is less than α, the level of significance chosen, the alternative hypothesis is true [194]. For a digit/sample combination, the ideal and experimental error rates can be matched as a pair. These error rates are independent of each other and are dependent paired t-test can be used. To perform the paired test, it is required that both ideal and experimental errors have a normal distribution.

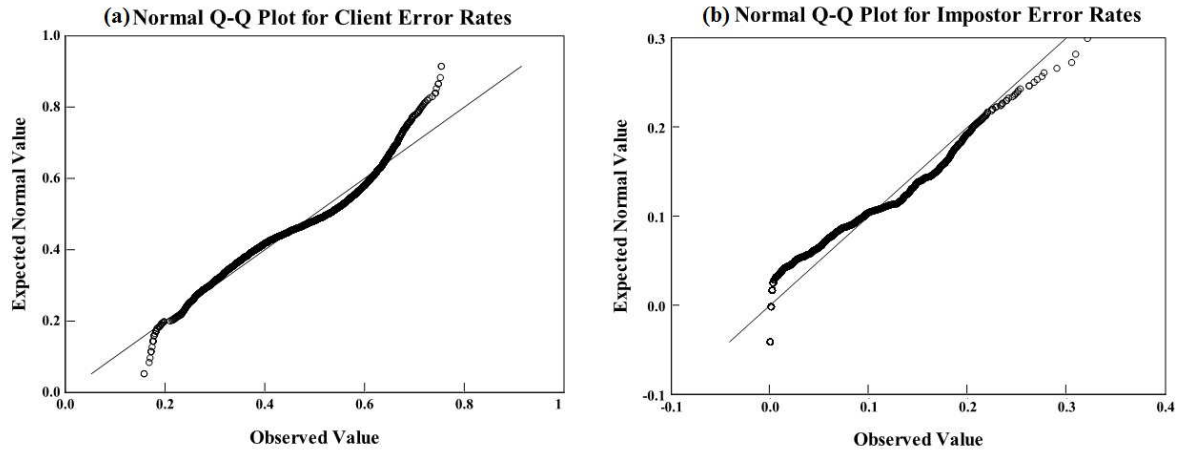

Figure 4.23 Normal Q-Q plots for the Client and Impostor samples from *SET-2*

The purpose of a normality analysis is to check whether the observed data supports the null hypothesis that the underlying probability density function is Normal. The Normal Q-Q plot provides a graphical representation for determining the normality. The normality assumptions in the strict sense of the word are not fulfilled for the error rates of this test data (fig. 4.24). Nevertheless, the results obtained by t-tests can be accepted, if the significance level is far away from the critical value (0.05 for the 95% confidence interval). The tests are performed to determine the dependence between decisions for each of the fusion technique, i.e., multi-instance, multi-sample and multi-instance & multi-sample fusion schemes.

4.7.1 Multi-Instance Fusion

The expressions for the verification error rates of multi-instance fusion (4.1) are presented with the assumption of equal FAR and equal FRR for each instance for the purpose of simplicity. These error rates for multiple instances, in general, can be different (e.g., different error rates for isolated digits represented in fig. 3.4(b)). Therefore the assumption of equal error rates can be relaxed for more complicated and exact formulae with $\alpha_{(i,1)}$ & $\rho_{(i,1)}$ as FAR and FRR for instance *'i'* ($i = 1, 2, 3, ...n$) with statistical independence between the decisions [23]. The expressions are given as:

$$\alpha_{Ideal}(n,1) = \alpha_{1,1}\alpha_{2,1}\alpha_{3,1}.....\alpha_{n,1} \tag{4.2}$$

$$\rho_{Ideal}(n,1) = \rho_{1,1} + (1-\rho_{1,1})\rho_{2,1} + ... + (1-\rho_{1,1})(1-\rho_{2,1})....(1-\rho_{n-1,1})\rho_{n,1} \tag{4.3}$$

The difference between 'ideal error rates' [1](4.2 & 4.3) and the experimental error rates is analysed using the paired t-test. The results presented in table 4.3 shows a significant difference in mean ideal error rates and experimental error rates. For false rejections, the fusion of two digits has the mean and standard deviation of 0.396 and 0.189 for ideal errors whereas a mean of 0.322 and standard deviation of 0.121 for experimental errors. As the corresponding p-value value is far less than 0.05, the null hypothesis is rejected with high confidence. The tests for fusion of additional digits have also shown that the statistical difference between ideal and experimental error rates is not zero. This analysis however does

[1] *'Ideal'* implies theoretically calculated error rates and not the lowest achievable error rates

Table 4.4 Paired t-Test: Paired Ideal and Experimental Error Rates for Means of Multi-instance fusion without repetitive samples

	Mean	$S.D.$[1]	95% Confidence Interval - Difference		t[2]	df[3]	p[4]
			Lower	Upper			
FRR : 2 Digits	0.073	0.064	0.066	0.080	20.57	323	7.25E-56
3 Digits	0.132	0.104	0.124	0.139	34.93	755	7.6E-160
4 Digits	0.170	0.127	0.162	0.177	45.11	1133	3.2E-255
5 Digits	0.192	0.139	0.184	0.200	46.69	1133	2.9E-266
6 Digits	0.204	0.143	0.194	0.214	39.19	755	3.8E-184
7 Digits	0.206	0.143	0.191	0.222	26.00	323	3.33E-81
8 Digits	0.204	0.140	0.173	0.235	13.10	80	1.35E-21
FAR : 2 Digits	-0.065	0.048	-0.070	-0.059	-24.05	323	5.76E-74
3 Digits	-0.072	0.060	-0.077	-0.068	-32.95	755	3.1E-148
4 Digits	-0.065	0.060	-0.068	-0.061	-36.58	1133	4.5E-194
5 Digits	-0.055	0.055	-0.059	-0.052	-33.64	1133	1.5E-172
6 Digits	-0.047	0.050	-0.050	-0.043	-25.50	755	6.1E-104
7 Digits	-0.039	0.045	-0.044	-0.034	-15.61	323	2.52E-41
8 Digits	-0.033	0.040	-0.042	-0.024	-7.32	80	1.73E-10

1: Standard deviation 2: Paired sample *t*-test

3: Degrees of freedom 4: *p* value. Significance criterion set at $p \leq 0.05$

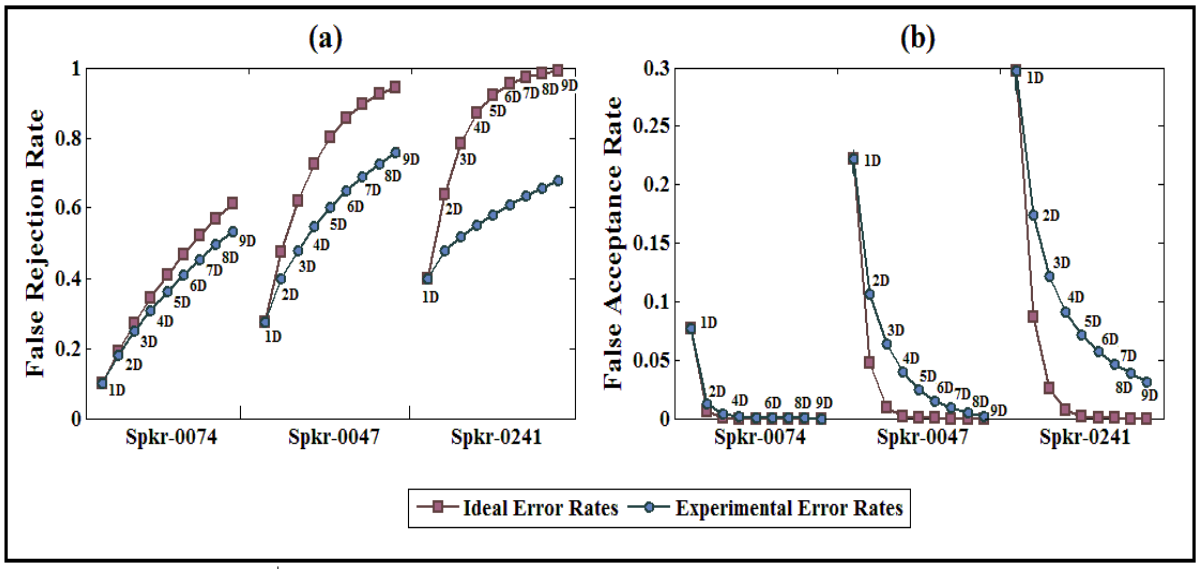

Figure 4.24 Comparison of Ideal and Experimental Error Rates for Multi-instance fusion schemes for three speakers (0074, 0047 & 0241) from *SET-2*

not determine if the fusion of independent decisions (ideal errors) is better than the statistically dependent decisions (experimental errors).

The results of the paired-t test (table 4.4) for multi-instance fusion errors are supported in fig. 4.25 where the mean ideal and experimental error rates are plotted for three speakers from *SET-2*. The experimental FRRs for digit combinations are lower than the ideal FRRs and experimental FARs are higher than ideal FARs. With the increase in digit used for fusion, the difference between the errors (ideal and experimental) increases. As false acceptance rate decreases with increase in digits, the ideal error rates can soon reach zero, because of limited tests or low base FARs, thus the difference in errors could not be represented accurately in figure 4.25(b).

4.7.2 Multi-Sample Fusion

The error rates for the multi-sample fusion presented in (4.1) assumes that the FAR and FRR are equal for each of the '*m*' repeated samples. This assumption is valid when the repeated samples are randomly selected from the remaining set. Nevertheless, when the subsequent sample is an adaptation of pervious sample, the error rates for tests performed on these samples can be different. And so the assumption of equal error rates can be relaxed for more complicated and exact formulae with $\alpha_{(1,i)}$ & $\rho_{(1,i)}$ as FAR and FRR for sample *'i'*

$(i = 1, 2, 3, ...m)$ with statistical independence between the decisions [23]. The expressions are given as:

$$\rho_{Ideal}(1, m) = \rho_{1,1}\rho_{1,2}\rho_{1,3}.....\rho_{1,n} \tag{4.4}$$

$$\alpha_{Ideal}(1, m) = \alpha_{1,1} + (1-\alpha_{1,1})\alpha_{1,2} + ... + (1-\alpha_{1,1})(1-\alpha_{1,2})....(1-\alpha_{1,m-1})\alpha_{1,m} \tag{4.5}$$

Table 4.5 presents the paired t-test results of the ideal and experimental error rates for fusion of random samples. The test significance for the false rejection rate and false

Table 4.5 Paired t-Test: Paired Ideal and Experimental Error Rates for Means of Multi-sample Fusion for individual digits

	Mean	$S.D^1$	95% Confidence Interval - Difference		t^2	df^3	p^4
			Lower	Upper			
FRR: 2 Samples	-0.002	0.046	-0.013	0.008	-0.467	80	.642
3 Samples	0.001	0.023	-0.004	0.006	0.476	80	.636
4 Samples	0.001	0.009	-0.001	0.003	0.770	80	.443
5 Samples	-0.001	0.005	-0.002	0.000	-1.703	80	.092
FAR: 2 Samples	-0.013	0.112	-0.038	0.012	-1.029	80	.307
3 Samples	-0.019	0.132	-0.048	0.011	-1.271	80	.207
4 Samples	-0.021	0.139	-0.051	0.010	-1.333	80	.186
5 Samples	-0.023	0.141	-0.055	0.008	-1.496	80	.139

1: Standard deviation 2: Paired sample *t*-test

3: Degrees of freedom 4: *p* value. Significance criterion set at $p \leq 0.05$

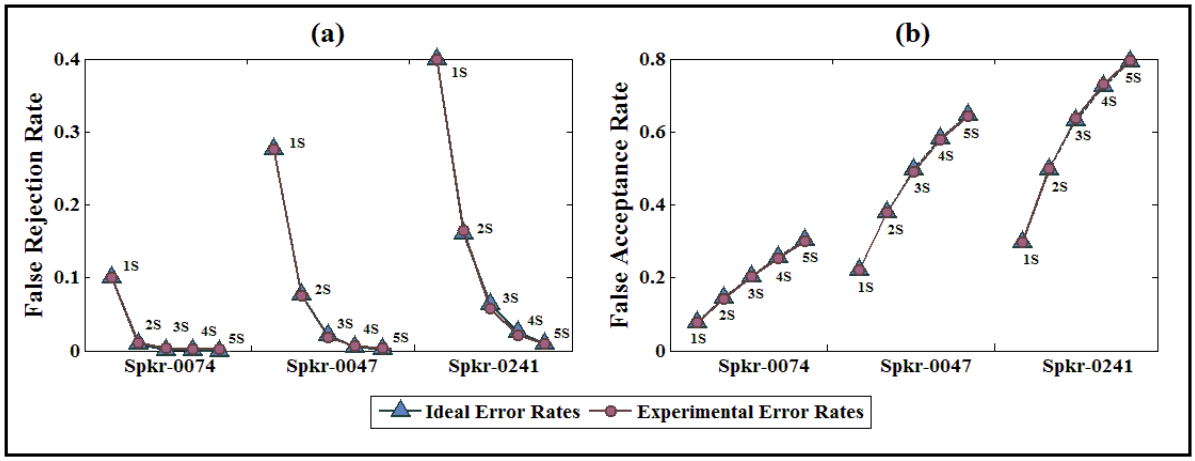

Figure 4.25 Comparison of Ideal and Experimental Error Rates for Multi-sample fusion schemes for three speakers (0074, 0047 & 0241) from *SET-2*

acceptance rate of multi-sample fusion is higher than 0.05 and so the null hypothesis that the difference between the ideal and experimental error rates is zero can be strongly accepted. Since the acceptance is, so strong for most of the cases it can be concluded that combinations for multiple samples have the same means for ideal and experimental error rates. This analysis of paired t-test is confirmed using the false rejection and false acceptance rates plotted in fig. 4.25(a) and (b) respectively. This difference between ideal and experimental errors depends on the performance of base digit models.

As the multi-sample fusion of adaptive samples is analysed using data from *SET-1*, the statistics of paired t-test for fusion of random and adaptive samples are shown in table A.2. The p value for the combinations of random sample is greater than 0.005, as is the case with data from *SET-2*, the null hypothesises accepted. For adaptive samples, the significance is less than 0.05 and the null hypothesises, that difference between ideal and experimental error rates is zero, is rejected (table 5.). The ideal errors for adaptive samples are higher/lower than the experimental error rates.

4.7.3 Multi-Instance and Multi-Sample Fusion

The false rejection rate for fusion of '*m*' samples (combined using *'AND'* rule) with individual FRR of $\rho_{(1,m)}$ for the '*1*st' instance is given as:

$$\rho_{Sample}(1,m) = \rho_{1,1}\rho_{1,2}\rho_{1,3}....\rho_{1,m} = \prod_{j=1}^{m}\rho_{1,j}$$

Whereas, the FRR for the fusion of '*n*' instances (combined using *'OR'* rule) with no repeated samples is given as:

$$\rho_{\text{Instance}}(n,1) = \rho_{1,1} + (1-\rho_{1,1})\rho_{2,1} + ... + (1-\rho_{1,1})(1-\rho_{2,1})..(1-\rho_{n-1,1})\,p_{n,1}$$

The total false rejection rate for the combination of '*n*' instances with '*m*' samples allowed at each instance level is obtained by using the above two equations.

$$\rho_{Ideal}(n,m) = \prod_{j=1}^{m}\rho_{1,j} + \left(1-\prod_{j=1}^{m}\rho_{1,j}\right)\prod_{j=1}^{m}\rho_{2,j} + ... + \left(1-\prod_{j=1}^{m}\rho_{1,j}\right)...\left(1-\prod_{j=1}^{m}\rho_{n-1,j}\right)\prod_{j=1}^{m}\rho_{n,j}$$

$$= 1-\prod_{i=1}^{n}\left(1-\prod_{j=1}^{m}\rho_{i,j}\right) \qquad (4.6)$$

Similarly, the false acceptance rate for the *'OR'* fusion of *'m'* samples and *'AND'* fusion of *'n'* instances with individual FARs of $\alpha_{(1,m)}$ and $\alpha_{(n,1)}$ respectively is given as:

$$\alpha_{Sample}(1,m) = \alpha_{1,1} + (1-\alpha_{1,1})\alpha_{1,2} + (1-\alpha_{1,1})(1-\alpha_{1,2})\alpha_{1,3} + ... + (1-\alpha_{1,1})...(1-\alpha_{1,m-1})\alpha_{1,m}$$

$$\approx \sum_{i=1}^{m}\alpha_{1,i} \quad (\alpha_{1,i} << 1)$$

$$\alpha_{\text{Instance}}(n,1) = \alpha_{1,1}\alpha_{2,1}\alpha_{3,1}....\alpha_{n,1}$$

The false acceptance rate for multi-instance and multi-sample fusion with $\alpha_{i,j}$ as FAR for 'i^{th}' instance and 'j^{th}' *sample* $(i = 1,2,3,...n;\ j = 1,2,3...m)$ with statistical independence between the decisions is given as:

$$\alpha_{Ideal}(n,m) \approx \sum_{i=1}^{m}\alpha_{1,i} * \sum_{i=1}^{m}\alpha_{2,i} * ..... * \sum_{i=1}^{m}\alpha_{n,i}$$

$$\approx \prod_{i=1}^{n}\left(\sum_{j=1}^{m}\alpha_{i,j}\right) \qquad (4.7)$$

As the fusion performance depends on base classifier errors, the proposed fusion is analysed for individual speakers with different performances in fig. 26. The experimental FRRs for digit combinations are lower than the ideal FRRs. With the increase in digits used for fusion, the difference between the ideal and experimental errors increases. For each digit combination, the difference decreases with increase in samples used for fusion (Also shown in fig. 4.35(a) for the mean false rejection rates across *SET-2*). For strong classifiers, less

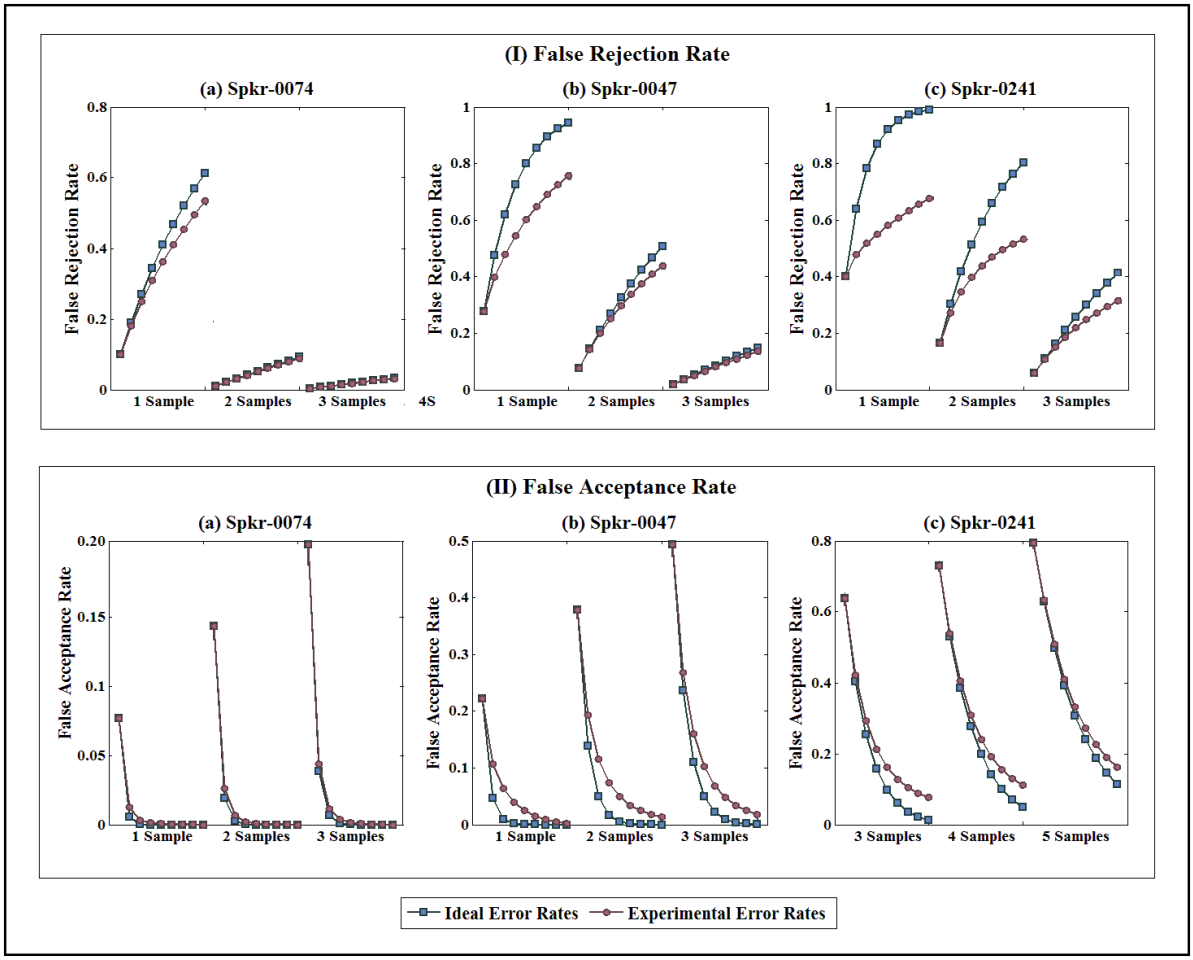

Figure 4.26 Mean Ideal and Experimental Error Rates for the proposed fusion schemes for (I) False Rejection Rate and (II) False Acceptance Rate for three speakers from SET-2 (a) Spkr-0074, (b) Spkr-0047 and (c) Spkr-0241

number of samples are required to reduce the difference between ideal and experimental FRR to zero (in fig.4.26 (I) fusion of four-samples with multiple digits has the same ideal and experimental FRR). The number of samples required for weak classifiers can be more depending on the error rates for base classifiers (in fig. 4.26 (II) for equal ideal and experimental errors more than five samples for each instance are required).

The experimental FARs is higher compared to ideal FARs for digit combinations. This difference decreases with increase in digits used for fusion. As false acceptance rate decreases with increase in digits, the ideal error rates soon reach zero value, because of limited tests or low base FARs. Thus, the difference in errors is represented for three-five samples for Spkr-0241 in fig. 4.26 II(c). The difference between mean ideal FARs and mean experimental FARs decreases with an increase in the samples used for fusion (Also shown in fig. 4.35(b) for the mean false acceptance rates across *SET-2*). For strong classifiers, less

number of multiple samples is required for zero difference between the ideal and experimental errors (fig. 4.26.I (a)) whereas comparably more number of samples may be required for weak classifiers (fig. 4.26.II (c)).

In most cases, the increase in samples used for fusion can result in catastrophic results and so the estimation of fusion error rates for such combinations might not be of significance. As a result it is important to explore the reasons for the difference between the ideal and experimental error rates. This difference in ideal and experimental mean error rates for the proposed fusion is because of either multi-instance or multi-sample fusion schemes. The general reasons for this difference can be explained because of

1. the statistical dependence (correlation) between classifier decisions resulting in error rates that are larger or smaller than the ideal values obtained under independence assumption [37, 195].
2. the correlation between the input data presented at each classifier even though the text is different [196].

In addition to the above reasons, if the errors are estimated using the base error rates of tune dataset, the mismatch conditions between the tune/development and test dataset conditions can result in difference between error rates. The mismatch conditions could be because of the handset or nearby sources of noise and differing acoustics or even the mismatch introduced by the speakers themselves. A number of techniques have been proposed to compensate for various aspects of session variability. Techniques such as feature warping [112], mapping [113] has been used to produce more robust features and score compensation techniques such as H-and T-Norm has also been proposed. In [197], an efficient model training procedure is proposed for Gaussian mixture modelling to perform the optimization of speaker model and session variables required for training.

In this dissertation, the correlation between the classifier decisions has been investigated for further refinement of the statistical analysis performed for the proposed fusion. In the next chapter, the effect of correlation modelling for the multi-instance and multi-sample fusion scheme is investigated by introducing the correlation modelling to the sequential decision fusion. The correlation between the decisions is modelled using Bahadur –Lazarsfeld expansion. The correlation modelling enables better tuning of parameters, '*n*' the

number of classifiers and '*m*' the number of attempts/samples, for minimising the difference between the ideal and experimental error rates.

4.8 Chapter Summary and Conclusion

This chapter provides the empirical evaluation of the proposed multi-instance and multi-sample fusion scheme (chapter 2). The decisions from multiple instances and samples are combined in sequence where an *'AND Rule'* is used for multi-instance fusion and an *'OR Rule'* for multi-sample fusion. This method of decision combination is shown to be effective in controlling the trade-off between the false rejections and false acceptances for verification. This architecture is evaluated for text dependent speaker verification using Hidden Markov Model (HMM) based digit dependent speaker models.

When the decisions are combined using sequential fusion, reliable final decision is made from early few verification decisions (minimum amount of test data is required). The experimental evaluation for sequential fusion has shown that better error rates are obtained by considering each digit in the string as a separate instance for verification rather than performing verification on entire digit-string. The sequential decision fusion of multiple instances (digits) is also shown to provide a trade-off between the performance and number of instances used for fusion.

The evaluation of the multi-instance and multi-sample fusion schemes was performed independently to find the significance of each fusion scheme on the proposed architecture. The sequential *'AND fusion'* of multiple instances reduces the FAR and increase the FRR compared to the base classifier performance. The reduction in FARs decrease with each progressive addition of an instance. The increase in FRR also decreases with increase in the number of instances used for fusion. The results for sequential *'OR fusion'* of multiple samples are complementary to the sequential *'AND fusion'*. The FRRs and FARs for multi-sample fusion are lower and higher than the base classifier performances respectively. The decrease (FRR) and increase (FAR) in fusion error rates was reduced with increase in samples used for fusion.

The proposed method of fusing multiple instances and multiple samples is observed to reduce both the verification error rates simultaneously. The reduction in FRR and FAR for the fusion is achieved irrespective of the base classifier errors but the improvement in fusion performance for different classifiers is dependent on the base performances. The tuning of the

parameters, '*n*' classifiers and '*m*' attempts/samples had the potential to improve the performance of weaker classifiers. This analysis for the proposed fusion is examined for the variations in datasets, thresholds and models used for verification. It is demonstrated that performance for the combination of decisions using *'Equal False Rejection Rate'* as threshold selection criteria outperforms the fusion of decisions obtained using *'Equal False Acceptance Rate'* and *'Equal Error Rates'* criteria.

The experimental results provide the evidence for the evaluation of the equations that are used for the estimation of verification error rates. In general, the verification system that is properly tuned performs reasonably well for the test data (unknown data) when the tune and test datasets are assumed to be of similar performances. It is therefore possible to estimate the fusion error rates of test dataset using the tune dataset base classifier error rates and the developed equations. To justify the possibility of error rate estimation, a paired p-test is performed for the theoretical or ideal error rates (calculated using the formulae) and the experimental error rates. It is observed that the mean error rates (ideal and experimental) for the fusion of multiple instances are significantly different whereas the mean values are similar for multi-sample fusion. The experimental FRRs for digit combinations are lower than the ideal FRRs and experimental FARs are higher than ideal FARs. With the increase in digit used for fusion, the difference between the errors (ideal and experimental) increases.

The difference in the mean ideal and experimental error rates for the proposed fusion is lowered by increasing the number of samples used for fusion. The base classifier performance determines the number of samples required for the fusion process to obtain equal ideal and experimental values. For classifiers with high error rates, if the number of samples is high the fusion performance is observed to be lower than the base classifiers. The reasons for the difference in the error rates are significant for fusion performance estimation and thus the formulae for verification errors are modified to accommodate the modelling of these factors.

The formulae presented are developed with the assumption of statistical independence between the classifier decisions. The difference in ideal and experimental error rates might be because of the statistical dependence between the classifier decisions. The next chapter presents the analytical and experimental evaluation of the formula for the proposed fusion scheme using correlation modelling. It is evident from the empirical evaluation presented in this chapter that superior fusion performance can be obtained despite the seemingly ideal assumption that base classifiers make uncorrelated decisions.

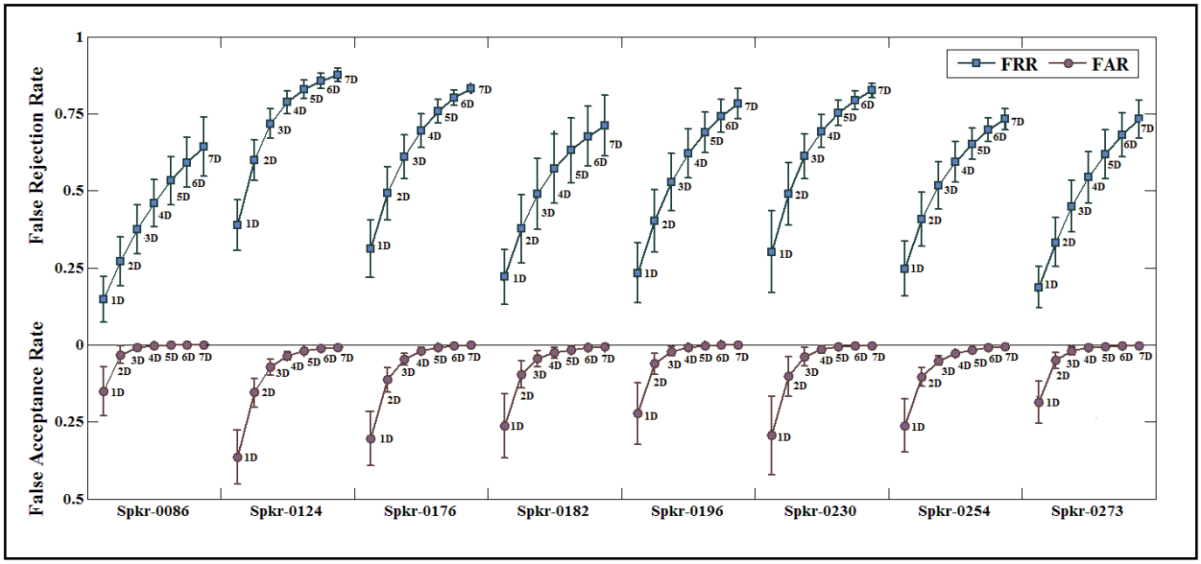

Figure 4.27 Verification error rates for Multi-instance fusion of speakers from *SET-1*

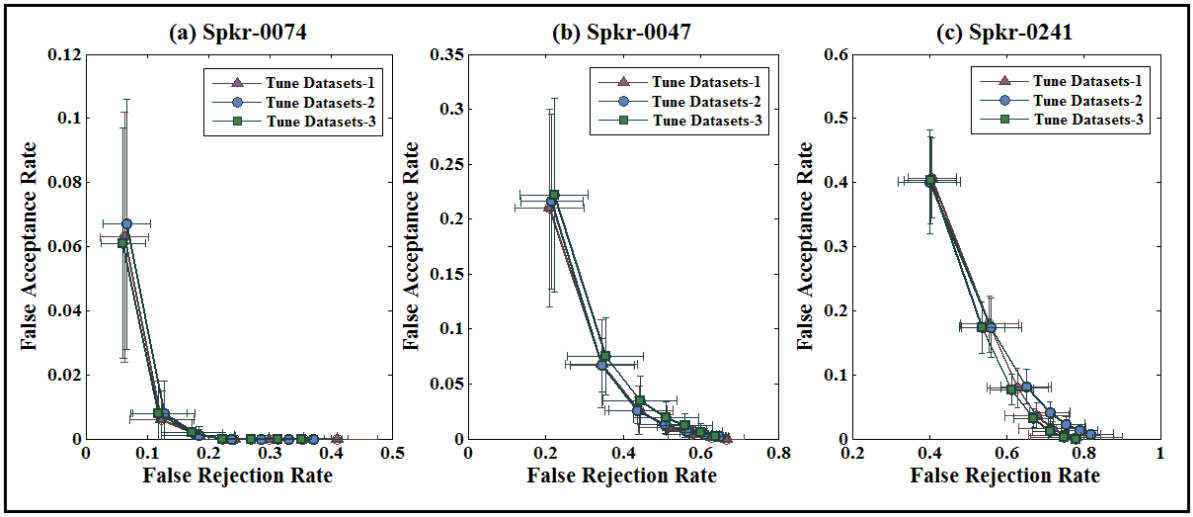

Figure 4.28 Multi-instance error rates for different datasets with data overlap for three speakers in *SET-1* (a) Spkr-0074, (b) Spkr-0047 and (c) Spkr-0241

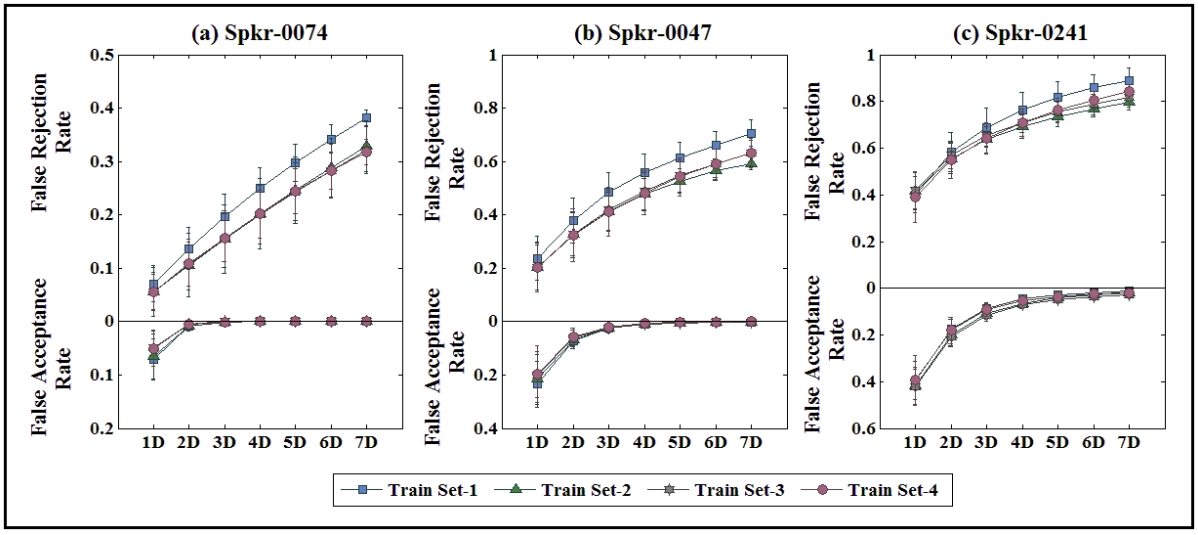

Figure 4.29 Verification error rates for multi-instance fusion of datasets tested on different training models for three speakers (a) Spkr-0074, (b) Spkr-0047 and (c) Spkr-0241 from *SET-1*

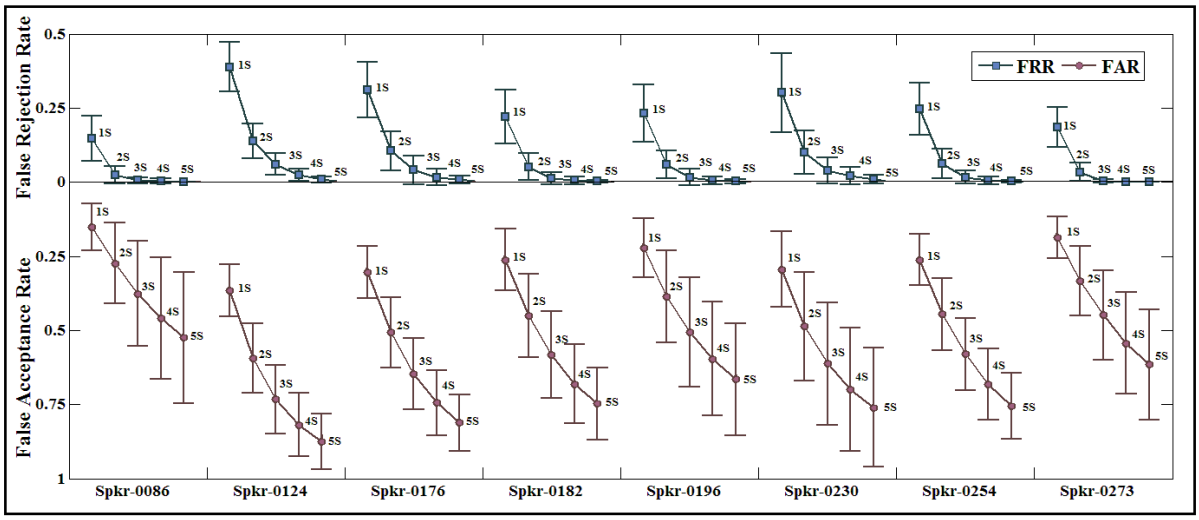

Figure 4.30 Verification error rates for multi-sample fusion of five randomly repeated digit samples for speakers from *SET-1*

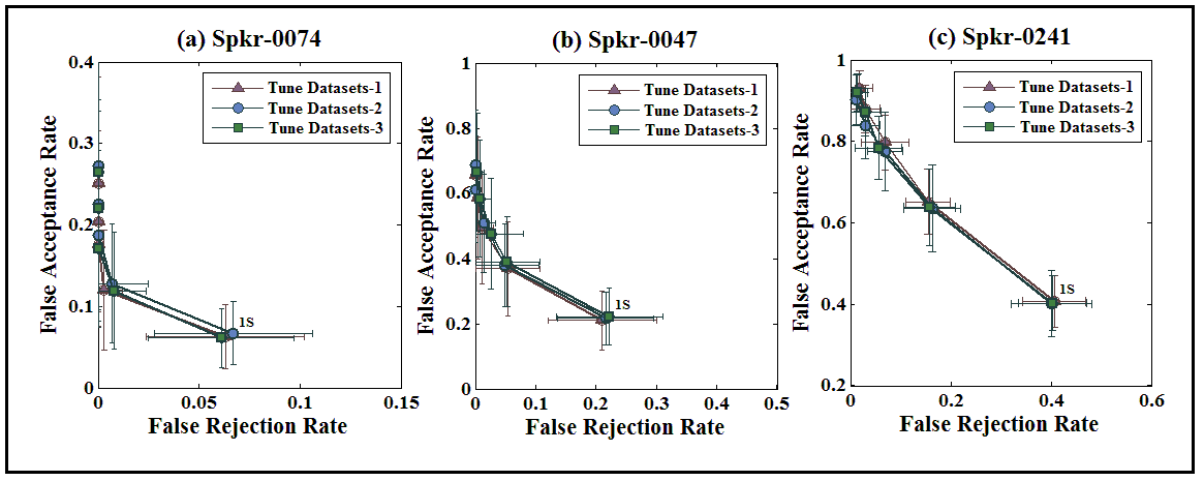

Figure 4.31 Multi-sample error rates for different datasets with data overlap for three speakers in *SET-1* (a) Spkr-0074, (b) Spkr-0047 and (c) Spkr-0241

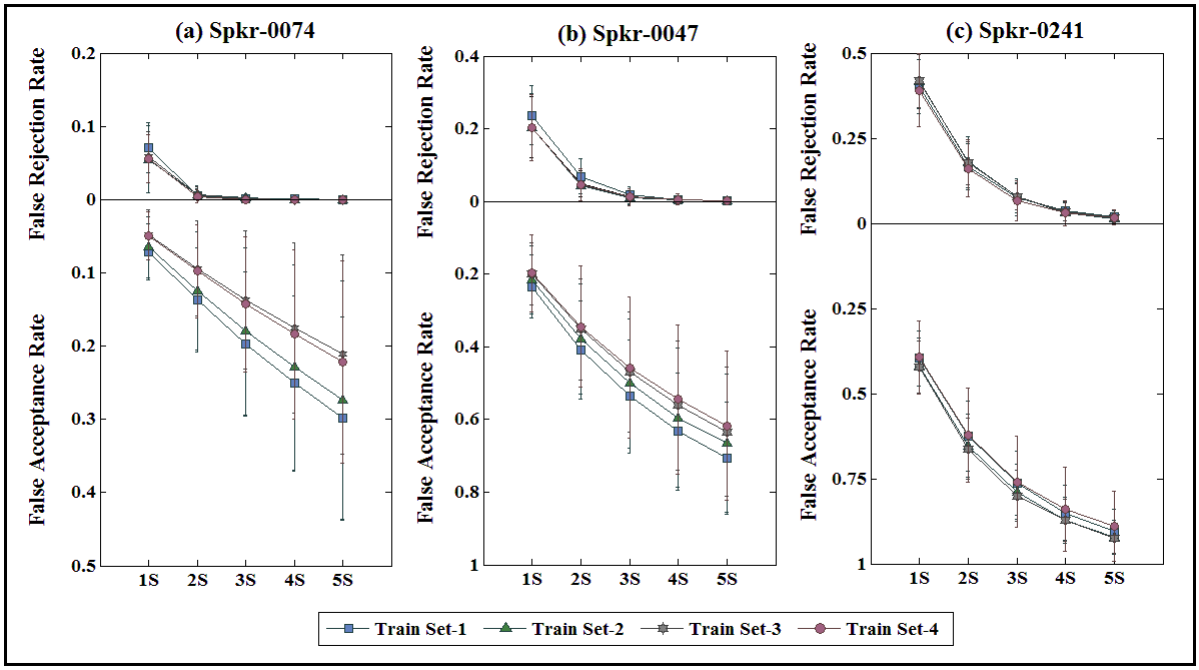

Figure 4.32 Verification error rates for multi-sample fusion of datasets tested on different training models for three speakers (a) Spkr-0074, (b) Spkr-0047 and (c) Spkr-0241 from *SET-1*

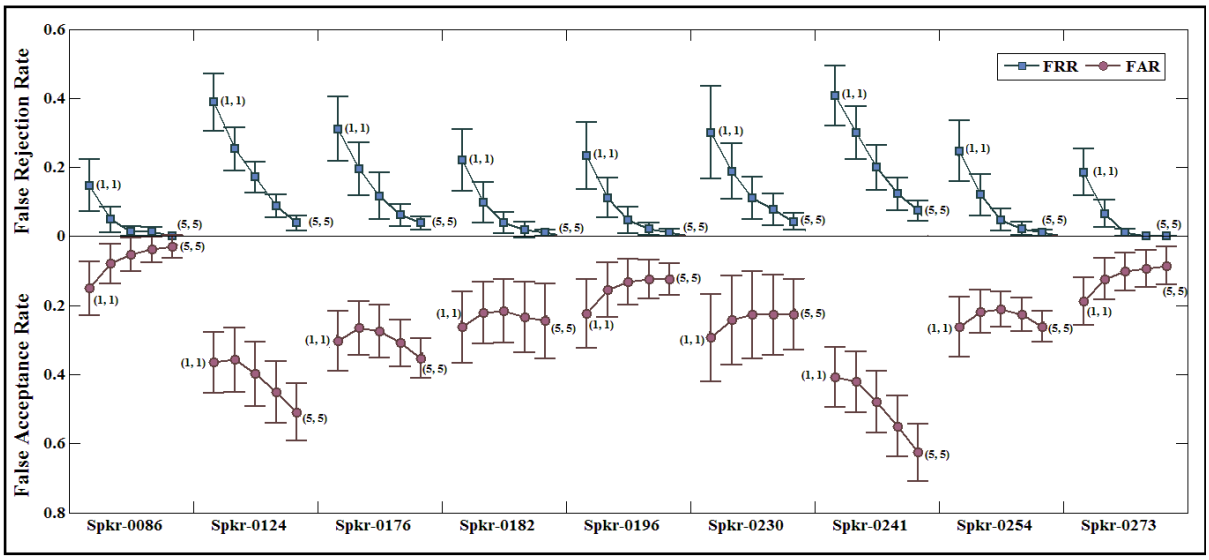

Figure 4.33 Verification error rates for integration of multi-instance and multi-sample fusion of speakers from *SET-1*

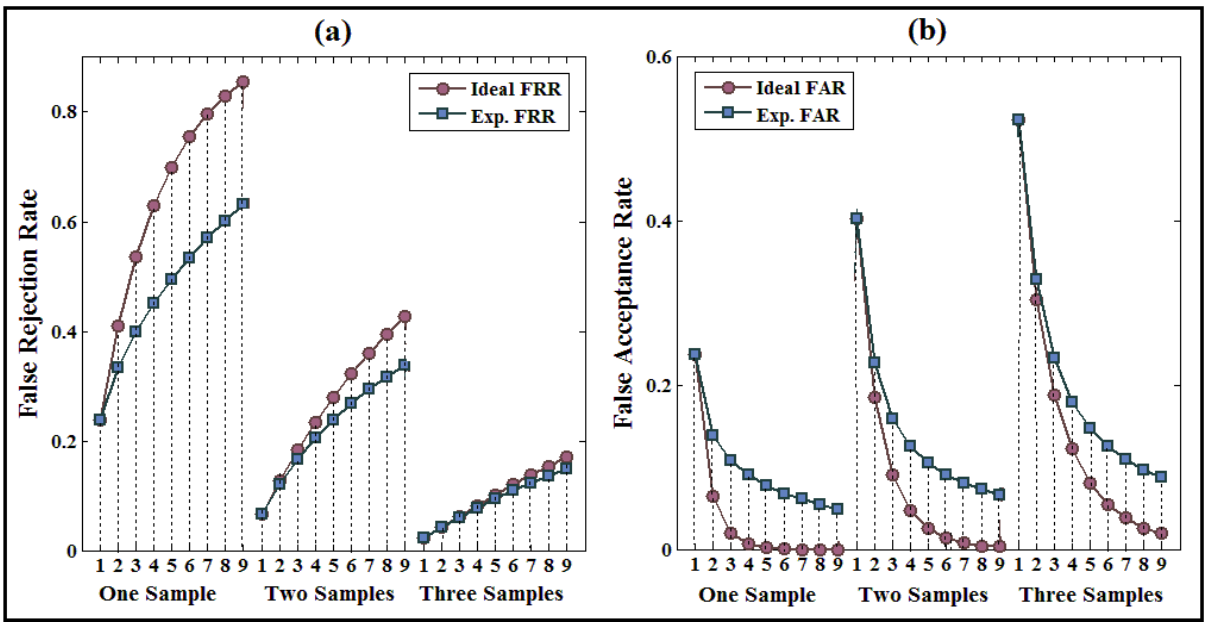

Figure 4.34 Mean Ideal and Experimental Error Rates for the proposed fusion schemes (a) False Rejection Rate and (b) False Acceptance Rate

Chapter 5
Modelling of Statistical Dependence between decisions for Proposed Fusion Scheme

5.1 Introduction

Several studies have shown that the performance of a single-biometric verification system can be improved by uni-modal fusion, i.e., the combination of several verification strategies applied on the data from multiple sources. Even greater improvement in verification performance is expected through the combination of statistically independent information from multiple biometric sources [15]. With some exceptions, the analysis of methods to combine classifiers assume independence on some level from either input data or outputs of the classifiers themselves (e.g., [198]). Many researchers have investigated whether fusion of independent matchers results in better performance compared to fusion of dependent matchers [48, 199]. Tulyakov et al. [200] have shown both theoretically and experimentally that utilizing independence provides better approximation of score density functions and results in performance improvement. Kittler et al. [16] also proposed a theoretical framework for combination using product rule at the decision-level with implicit assumption of classifier independence.

Although the assumption of statistical independence often seems unrealistic in many situations, it provides an adequate and workable approximation of the reality which may be more complex. Example is the Gaussian assumption frequently made even in situations where class distributions patently do not obey the exponential law. This assumption simplifies the understanding and yields acceptable results [16] and routinely used in practise. The analysis is shown to provide a plausible theoretical underpinning of the combination rules and thereby draws attention to the underlying assumptions behind these schemes that the users may not be aware. Similarly, the assumption of independence holds good for combination of different biometric characteristics ([16, 200]) which may not be true for general multibiometric system [34]. For example, multiple instances or samples from the same biometric characteristic, with or without similar feature sets, can be dependent because of the similarity in biometric data.

Domingos and Pazanni [201] have suggested that scenarios exist in which a classifier that assumes conditional independence is optimal even for highly correlated data. The correlation or lack of independence, in general, have some influence on classifier combination [122]. The level at which fusion is performed also has an effect on the level of dependence, i.e., moderate correlation has a negligible impact on accuracy when decisions from relatively accurate verifiers are combined rather than raw features [101]. The correlation between the classifiers can be incorporated into the fusion scheme [34] to obtain a significant performance improvement when compared with a simplified fusion scheme that is based on the independence assumption. There have been claims of marginal improvement (e.g.,[202]) in the fusion performance when correlation is considered but a systematic analysis of this problem has not been presented thus far.

The improvement in speaker verification performance for incorporation of correlation at various levels was investigated. Farrell et al. [88] analysed correlation using *phrase-level scores* obtained from modelling methods such as neural tree network (NTN), Gaussian mixture model (GMM), Hidden Markov Model (HMM), and dynamic time warping (DTW). This analysis, however, deals with correlation between errors of these modelling approaches. The results show that the best combination performance is obtained by combining the models with least correlation between errors rather than models with the best performance. Similar analysis has shown that 'negatively correlated' errors results in lower error rates than those that make uncorrelated errors [203]. Analysis on the correlation between successive *feature vectors* for improvement in speaker verification is proposed using sequential probability ratio test (SPRT) [153]. The correlation of *speaker differences* between a pair of digits (multiple instances) is analysed in [204]. It is found that the correlation of speaker differences (two different voices) between spoken digits is large comparatively except for few pairs. The experiments show that the digits can be divided into two groups (1, 6 and 8) and (2, 3, 4, 5, 7, 9 and 0), i.e., with and without devocalized vowels respectively. The investigation also shows that the correlation is not influenced by the speaker differences of silence parts in speech. The correlation of speech sounds produced by the same speaker is analysed by [205]. Based on this analysis a recognition algorithm is developed that classifies different sounds produced by the same speaker jointly. This intra-class correlation between the words (instances or multiple samples) from the same speaker can affect fusion performance.

The sequential decision fusion (multi-instance and multi-sample) system proposed in the dissertation is analytically shown to improve performance and allow a controlled trade-off

between false rejection rate (FRR) and false acceptance rate (FAR) when the classifier decisions are assumed to be statistically independent. However, it is concluded from previous chapter that the classifier decisions used for fusion may not be necessarily independent and so there exists a significant difference between ideal and experimental verification error rates. It is therefore important to analyse the effect of dependence between decisions of multiple sources of information from a single biometric characteristic. A statistical analysis of the problem of fusing independent classifier decisions has been addressed in the context of writer verification from different handwritten words in [196]. The analysis, however, did not consider the modelling of correlation between multiple samples combined in a sequential scheme. This chapter thus investigates the correlation modelling techniques for multi-instance, multi-sample and proposed architectures. The expressions developed for verification error rates are modified to incorporate correlation between classifier decisions.

Section 5.2 provides the theoretical analyses on modelling the statistical dependence between classifier decisions for multi-instance and multi-sample fusion schemes. The expressions are experimentally evaluated by considering the proposed architecture for text-dependent speaker verification using HMM based digit dependent speaker models. The analysis on *'favourable/unfavourable'* dependence between *'n'* instances and *'m'* samples is presented in section 5.3. This chapter also provides the empirical results for determining the order of correlation required for accurate estimation of error rates (section 5.4). The last section presents the analysis for estimation of verification error rates for unseen data from known parameters, such as the number of instances and number of samples, variance in correlation modelling and favourable digit combinations, fine-tuned using seen data.

5.2 Statistical dependence between decisions

OVERVIEW: Although the assumption of statistical independence simplifies the understanding of fusion and yields acceptable results, the assumption is unrealistic for multibiometrics from the same modality. The multiple instances or samples from same biometric characteristic can be statistically dependent because of the similarity in biometric data. Diversity can be analysed using probability distributions that represent the fusion error rates as a function of statistical dependence. Therefore, the dependence between classifier decisions, here, is analysed using class-conditional errors of the classifier fusion and class-conditional diversity values for 'AND' and 'OR' rules. The exact class-conditional error rates

for the fusion of correlated decisions are estimated using the full expansion of Bahadur-Lazarsfeld Expansion (BLE). This section provides the expression for BLE and the correction factor that includes the correlation coefficients between the decisions.

The expressions developed for the fusion of decisions from multiple instances, multiple samples, multiple instances and samples combined sequentially have shown to improve performance under the assumption of statistical independence (Section 2.7). As multiple instances/samples from the same biometric characteristic can be statistically dependent because of the similarity in biometric data, the expressions for the verification error rates are modified to incorporate correlation between the decisions. This section provides the theoretical analyses on modelling the statistical dependence between classifier decisions for multi-instance and multi-sample fusion schemes. The expressions are experimentally evaluated by considering the proposed architecture for text-dependent speaker verification using HMM based digit dependent speaker models.

The performance of a single–best classifier can be enhanced using the classifier fusion approach. Nevertheless, combining classifiers is expensive and is therefore significant to determine better performing classifier combinations. A well performing system can also be constructed out of individually weak classifiers that can be strong as a team [206]. However, combining identical classifiers does not gain any advantage over one such classifier. Therefore, diversity, also related to negative dependence, independence, orthogonality, complementarity, among the combining classifiers is a key issue. Diversity can be either beneficial or harmful as a set of dependent classifiers may result either better performance than the independent classifier combination performance or worse performance than the single worst classifier of the set depending on the difference. Therefore, understanding and measuring these differences (caused because of dependence between classifiers) in diversity is a significant issue in classifier combination.

The studies on diversity measures attempt to answer the few significant questions on diversity [39, 42].

1. How do we define and measure diversity?
2. How are the diversity measures related to the accuracy of the classifier combination?
3. Is there a best measure that is useful to describe minimum combination error?
4. How can we use the diversity measures to design the classifier combinations?

Kuncheva & Whitaker [42] analysed 10 different measures to define and quantify diversity of a group of classifiers using oracle representation that does not distinguish between types of outcome errors with identical misclassification costs. These measures that are pairwise or non-pairwise between and among binary classifiers are used to determine the strength of association between accuracy and each individual diversity measure [42]. For fusion schemes with performance better than independent classifier combination, the diversity measures are studied and a threshold is estimated to predict better performance than independent classifiers. This analysis, however, do not present a clear separation between diversity values for the more accurate classifier combinations and the less accurate combinations (than independent combinations) with equal classifiers misclassifications. The approach used to design classifier combinations also did not find a definite relation between diversity values and ensemble accuracy for a majority vote as it depends on the classifier combination strategy itself. The designed combinations for obtaining diversity such as AdaBoost classifiers could show a relation between the diversity and the classifier combination accuracy. The reason for these inconclusive results is the lack of understanding on the role of probability distributions in the analysis.

Gal-Or et al. [207] explored the relationship between diversity measures and classifier combination performance with incorporation of unequal misclassification costs that results in different and multiple performance measures. Bian et al. [208] studied the relationship between diversity and accuracy for homogeneous and heterogeneous ensembles. Venkataramani [39] investigated the questions related to diversity using joint probability distributions that enable the representation of optimal decision fusion accuracy as a function of statistical dependence and design classifier ensembles (using favourable joint probability distributions). This analysis also considers the class-conditional errors and class-conditional diversity values for *AND, OR* and Majority decision fusion rules. The relationship between the class-conditional error rates and class-conditional dependence can be modelled using Bahadur-Lazarsfeld Expansion (BLE). Kam et al. [209] addressed the problem of expressing the verification likelihood ratios, when local decisions are correlated, using the Bahadur-Lazarsfeld polynomials to form an expansion of the probability density functions in terms of normalized decisions and correlation coefficients of the normalized decisions.

The error rates when calculated with full expansion of Bahadur-Lazarsfeld Expansion (BLE) provide a more accurate estimation of errors. The expansion begins with the estimate of error rates under independence assumption, and this is multiplied by a correction factor.

The correction factor consists of a series of individual factors with correlations between multiple terms and a factor, such that when the full expansion is used, the exact error is computed. When the components of the vector *X* are discrete, the problem of estimating density becomes the problem of estimating probability *P(X)*. If the components of *X* are statistically independent, the problem is greatly simplified and is given as

$$P(X) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{1-x_i} \tag{5.1}$$

Where $p_i = P(x_i = 1)$ and $1 - p_i = P(x_i = 0)$. For correlated decisions, the Bahadur-Lazarsfeld expansion is used to express the density *P(X)* as [196]

$$P(X) = \prod_{i=1}^{n} (p_i^{x_i} (1 - p_i)^{1-x_i}) * \left[1 + \sum_{i<j} \gamma_{ij} z_i z_j + \sum_{i<j<k} \gamma_{ijk} z_i z_j z_k + ... \right] \tag{5.2}$$

Where γs are the correlation coefficients of the corresponding variables defined using z_i's, variables that are orthogonal with respect to the independence model with zero mean and unit variance

$$\gamma_{123...n} = \sum_X z_1 z_2 z_3 ..... z_n P(X) \text{ and } z_i = \frac{(x_i - p_i)}{\sqrt{p_i (1 - p_i)}} \tag{5.3}$$

The expansion 5.1 contains $2^n - 1$ coefficients, the *n* first-order probabilities p_i, the $\binom{n}{2}$ 2nd-order correlation coefficients γ_{ij}, the $\binom{n}{3}$ 2nd-order correlation coefficients γ_{ijk}, and so on. Specifically, correlation coefficients can be computed as:

- 2nd-order Correlation Coefficient:

$$\gamma_{ij} = \frac{E\left[\left(d_i - p_i \right) \left(d_j - p_j \right) \right]}{\sqrt{p_i p_j \left(1 - p_i \right) \left(1 - p_j \right)}} = \frac{E\left(d_i d_j \right) - p_i p_j}{\sqrt{p_i p_j \left(1 - p_i \right) \left(1 - p_j \right)}} \tag{5.4}$$

- 3rd-order Correlation Coefficient:

$$\gamma_{ijk} = \frac{E\left[\left(d_i - p_i\right)\left(d_j - p_j\right)\left(d_k - p_k\right)\right]}{\sqrt{p_i p_j p_k \left(1-p_i\right)\left(1-p_j\right)\left(1-p_k\right)}}$$

$$= \frac{E\left(d_i d_j d_k\right) - E\left(d_i d_j\right) p_k - E\left(d_i d_k\right) p_j - E\left(d_j d_k\right) p_i + 2 p_i p_j p_k}{\sqrt{p_i p_j p_k \left(1-p_i\right)\left(1-p_j\right)\left(1-p_k\right)}}$$

$$= \frac{E\left(d_i d_j d_k\right) - p_i p_j p_k}{\sqrt{p_i p_j p_k \left(1-p_i\right)\left(1-p_j\right)\left(1-p_k\right)}} - \sum_{i<j} \gamma_{ij} \sum_{k=1}^{3} \sum_{k\neq i, k\neq j} \sqrt{\frac{p_k}{\left(1-p_k\right)}} \tag{5.5}$$

- 4th-order Correlation Coefficient:

$$\gamma_{ijkl} = \frac{E\left[\left(d_i - p_i\right)\left(d_j - p_j\right)\left(d_k - p_k\right)\left(d_l - p_l\right)\right]}{\sqrt{p_i p_j p_k p_l \left(1-p_i\right)\left(1-p_j\right)\left(1-p_k\right)\left(1-p_l\right)}}$$

$$= \frac{E\left(d_i d_j d_k d_l\right) - p_i p_j p_k p_l}{\sqrt{p_i p_j p_k p_l \left(1-p_i\right)\left(1-p_j\right)\left(1-p_k\right)\left(1-p_l\right)}} + \sum_{i<j<k} \gamma_{ijk} \sum_{l=1}^{4} \sum_{l\neq i, l\neq j, l\neq k} \sqrt{\frac{p_l}{\left(1-p_l\right)}} - \sum_{i<j} \gamma_{ij} \sum_{k=1}^{4} \sum_{k\neq i, k\neq j} \sqrt{\frac{p_k}{\left(1-p_k\right)}} \tag{5.6}$$

- nth Order Correlation Coefficient:

$$\gamma_{ijk\ldots n} = \frac{E\left[\left(d_i - p_i\right)\left(d_j - p_j\right)\left(d_k - p_k\right).....\left(d_n - p_n\right)\right]}{\sqrt{p_i p_j p_k ....p_n \left(1-p_i\right)\left(1-p_j\right)\left(1-p_k\right).....\left(1-p_n\right)}} \tag{5.7}$$

The effects of correlation on classifier fusion have been investigated in the literature for sensor fusion and classifier decision fusion. Ali and Pazzani [203] discussed the relationship between error correlations and error reductions in the context of decision trees

and have shown that 'negatively correlated' errors results in lower error rates than those that make uncorrelated errors. Tumer and Ghosh [210] quantified the influence of the correlation between the classifiers on the error rate of multiple classifiers. They have shown that positively correlated classifiers reduced the added error only slightly, uncorrelated classifiers reduced the added error by a factor of 1/*n* and negatively correlated classifiers reduced the error even further. However, there is a limit on the largest absolute value of a negative pair wise correlation among *n* classifiers. Jacobs [211] reported that *'n'* dependent classifiers are worth as much as *'m'* independent classifiers where $m \leq n$. It is also shown that in some cases exact value for *'m'* can be given and in other cases, lower and upper bounds can be placed on *'m'*.

Sanchez et al. [3] showed theoretically and empirically that fusion of *'n'* instances/classifiers of a biometric trait reduces the system error by as much as 40% for statistically independent expert opinions. The use of multiple samples under independence assumption has been investigated by Kashi and Nelson [22] for signature verification and Jain et al. [11] for fingerprint verification (multiple impressions of the same finger). Cheung et al. [26] proposed a single-source multi-sample fusion approach for text-independent speaker verification. Here, the scores from claimant's utterances/samples with different weights are averaged and the resulting mean score is used for decision-making. The samples extracted for fusion irrespective of the modality are considered independent for all the above-explained evaluations. The assumption of independence between samples and/or instances from the same modality do not hold true and therefore the accurate estimation of fusion error rates requires the modelling of dependence between the decisions.

5.2.1 Fusion of *'n'* instances

The dependence between the decisions from multiple instances $d_i (i = 1, 2, 3 \ldots n)$ is estimated based on the Bahadur-Lazarsfeld Expansion (BLE) [196]. The expansion begins with the calculation of ideal error rates that are multiplied with a correction factor. The expressions for the ideal error rates (independent decisions) of multi-instance fusion schemes are given as

$$p^0_{Ideal} = \alpha_1 \alpha_2 \alpha_3 \ldots \alpha_n \quad \left(\alpha_i = FAR_i \text{ for instance } i\right) \tag{5.8}$$

$$\rho^{1}_{Ideal} = \beta_1\beta_2\beta_3...\beta_n \left(\beta_i = (1 - FRR_i) \text{ for instance } i \right) \tag{5.9}$$

Here ideal refers to the case of statistically independent decisions. Superscript '0' refers to FAR and '1' to FRR whereas subscripts identify the multiple instances in (5.8) & (5.9). α's and β 's are the base classifier FAR and (1-FRR). The equations to calculate the error rates for multi-instance fusion *with incorporation of correlation* between the decisions is given [196] as

$$p^{a}_{Est} = p^{a}_{Ideal}\left(1 + \sum_{i<j} \gamma^{a}_{ij} z^{a}_{i} z^{a}_{j} + \sum_{i<j<k} \gamma^{a}_{ijk} z^{a}_{i} z^{a}_{j} z^{a}_{k} + ...\right) \tag{5.10}$$

$$\text{Where } \gamma^{a}_{123...n} = E\left[z^{a}_{1} z^{a}_{2} ....z^{a}_{n} \right] \text{ and } \quad z^{0}_{i} = \frac{d_i - \alpha_i}{\sqrt{\alpha_i (1-\alpha_i)}}, \; z^{1}_{i} = \frac{d_i - \beta_i}{\sqrt{\beta_i (1-\beta_i)}} \tag{5.11}$$

Here, $\gamma^{a} (a = 0,1)$ are the correlation coefficients for true speaker and impostor decisions. They are defined using z_i variables that are orthogonal with respect to the independence model with zero mean and unit variance. Decisions d_i are 1 for client & 0 for impostors and so z_i^{0} are positive for incorrect impostor decisions and negative for correct ones. If two classifiers/instances are such that one is correct when the other is not and vice versa most of the time, these variables contribute to negative correlation. Client decisions are similarly handled with z_i^{1}. The magnitude and sign of the correlation, however, depends on the summation over all combinations. The expansion continues to third and higher-order decision correlations between classifiers.

For correlated decisions, the ideal error rates calculated using the (5.8) & (5.9) under independence assumption are different from experimental error rates or those predicted after applying correlation values in (5.10). The limitation of the analysis presented for sequential decision fusion (sections 2.7 & 4.8) is that the assumption of independent decisions from different instances may not be true for several words spoken by the same individual. Although these decisions are usually obtained using independent feature extraction and/or modelling procedures, the information content of each digit/word is not very different from others. For example, the phonemes involved between the digits may be the same and so the verification decisions from these digits can be correlated.

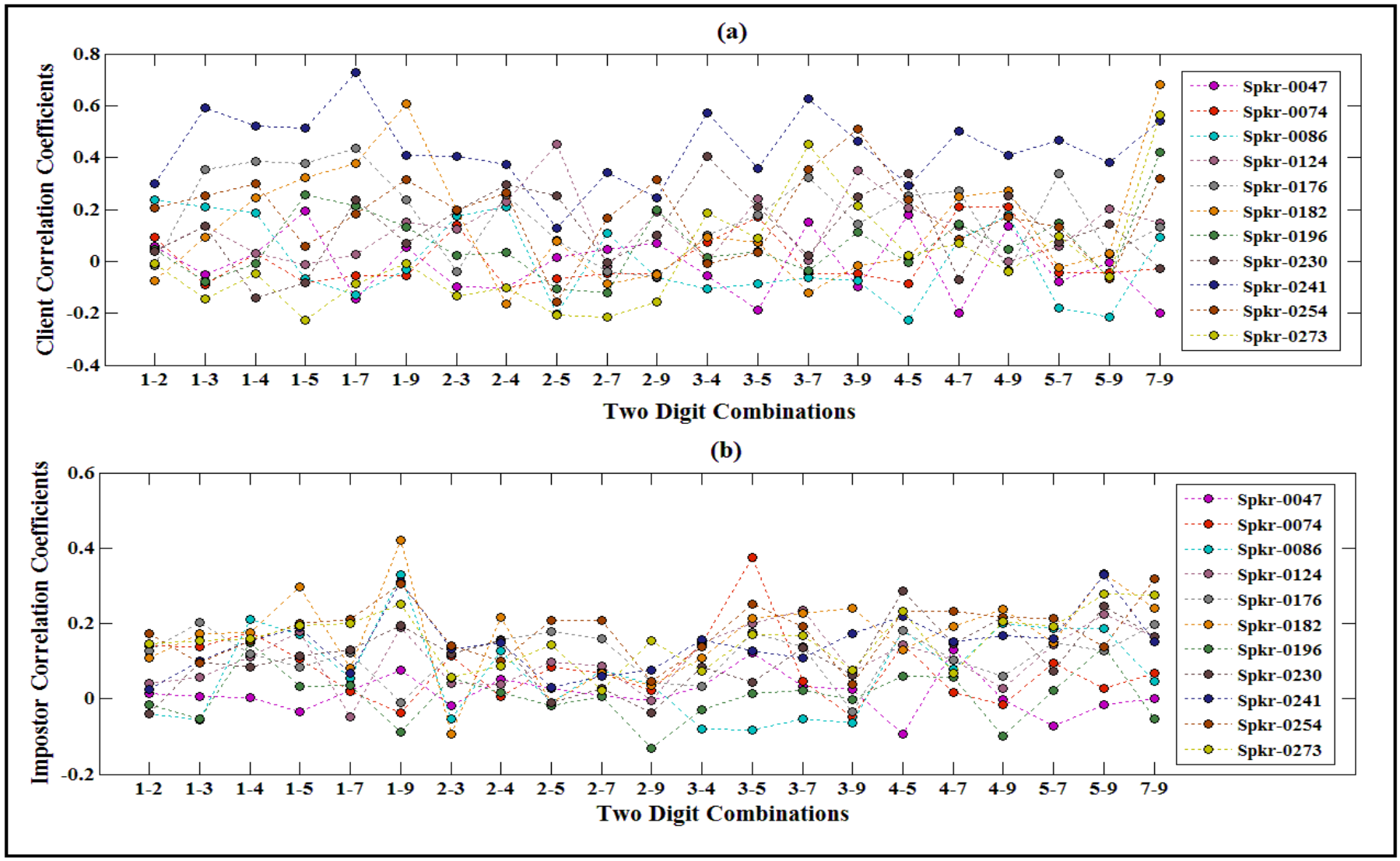

Figure 5.1 2nd-Order Correlation Coefficients for fusion of two digits from speakers of *SET-1* (a) Client Decisions and (b) Impostor Decisions

The error rates for the fusion of multiple instances calculated under dependence assumption are higher/lower than the experimental values (section 4.8). One of the expected reasons for this difference in error rates is the correlation between classifier decisions. To evaluate the effect of dependence between classifier decisions, the correlation coefficients are calculated using (5.2). The protocol used for this evaluation is explained in section 3.5.2. The term ***multi-instance fusion*** here is referred to as the ***combination of digits***. The fusion scheme in this section is evaluated for speech data from *SET-1*. The experimental values presented the pooled results for all speakers' test datasets.

Figure 5.1 shows the 2nd-order correlation coefficients between two-digit decisions for client and impostor decisions of *SET-1*. As the calculation of correlation coefficients depends on the base error rates (5.4-5.7), the fusion of instances/classifiers with different base performances is observed to have different correlation coefficients. Further, the correlation between the same two digits for multiple speakers is demonstrated to be different in both the sign and magnitude. Usually, the sign (positive or negative) of the correlation coefficient represents the direction of relationship where as the size/magnitude of the coefficient indicates the strength of relationship. Although the figure represents correlation values for two-digit combinations, it is generalised that the correlation values can be varied between different speakers for the same digit combination and between digit combinations for the same speaker.

In fig. 5.1, the speaker-dependent 2nd-order correlation coefficients for two-digit combination are represented. The mean correlation values for the 2nd-7th order coefficients are shown in the table 5.1. The correlation values (5.2) and the ideal error rates (5.8 & 5.9) are used to determine the predicted error rates - when calculated approximately are equal to experimental error rates - for each digit combination of a speaker. Table 5.1 presents the mean values for the ideal and experimental error rates for the combination of '*n*' digits (*n* =1, 2, 3...7). The *ideal false rejection rates* (FRR) are observed to be higher than the *experimental/predicted FRRs* whereas the *ideal false acceptance rates* (FAR) are lower than the *experimental/predicted FARs* for multi-instance fusion. This difference in error rates is due to the presence of statistical dependence (positive or negative correlation - table 5.1) between the decisions. The correlation coefficients for *impostor decisions* are demonstrated to be *positive* for the digit combinations. Here, the error rates for fusion of independent decisions are lower than dependent decisions. The *even order correlation coefficients*, i.e.,

Table 5.1 Mean ideal and experimental error rates with correlation coefficients (2nd-7th Order) for decisions from multiple instances

	Ideal Error Rates		Experimental Error Rates		Correlation Coefficients	
	FRR	FAR	FRR	FAR	FRR	FAR
Two Digits	$0.405^{\pm 0.20}$	$0.070^{\pm 0.07}$	$0.372^{\pm 0.18}$	$0.089^{\pm 0.07}$	$0.169^{\pm 0.2}$	$0.111^{\pm 0.10}$
Three Digits	$0.526^{\pm 0.22}$	$0.023^{\pm 0.02}$	$0.465^{\pm 0.19}$	$0.041^{\pm 0.04}$	$-0.069^{\pm 0.18}$	$0.031^{\pm 0.11}$
Four Digits	$0.614^{\pm 0.23}$	$0.008^{\pm 0.05}$	$0.535^{\pm 0.20}$	$0.021^{\pm 0.02}$	$0.115^{\pm 0.23}$	$0.044^{0.16}$
Five Digits	$0.680^{\pm 0.24}$	$0.003^{\pm 0.003}$	$0.589^{\pm 0.20}$	$0.012^{\pm 0.01}$	$-0.070^{\pm 0.24}$	$0.019^{\pm 0.16}$
Six Digits	$0.731^{\pm 0.24}$	$0.001^{\pm 0.001}$	$0.633^{\pm 0.19}$	$0.008^{\pm 0.005}$	$0.053^{\pm 0.27}$	$0.023^{\pm 0.17}$
Seven Digits	$0.770^{\pm 0.23}$	Ω [2]	$0.669^{\pm 0.19}$	$0.005^{\pm 0.005}$	$0.014^{\pm 0.26}$	$0.014^{\pm 0.27}$

correlations for fusion of two and four-digit combinations are *positive* whereas the *odd order correlation coefficients* for fusion of three and five-digit combinations are *negative*.

Although the sign of the correlation coefficients has been used in the literature to determine the dependence [73], this criterion alone may not be sufficient for better understanding of the role of dependence between decisions on performance of the proposed fusion. The analysis of '*favourable/unfavourable'* dependence (section 5.3) can be used to better determine if the fusion of correlated decisions results in lower/higher error rates compared to the fusion of independent decisions from multiple instances.

5.2.2 Fusion of '*m*' samples

The decisions from multiple samples are combined in the fusion process in order to obtain a reliable decision at an instance level. The Bahadur-Lazarsfeld Expansion is also applicable for the prediction of error rates *with incorporation of correlation* between the decisions [212] for multi-sample fusion. The expansion begins with the calculation of ideal

[2] Number of tests performed is limited and thus the error rates reaches zero.

error rates that are multiplied with a correction factor. The ideal error rates (under the assumption of statistical independence between decisions) for the *'OR fusion'* of multiple samples [32] are

$$p^0_{Ideal} = \delta_1\delta_2\delta_3...\delta_m \ \left(\delta_i = \left(1 - FAR_i\right) \text{for } i^{th} \text{ sample}\right) \tag{5.12}$$

$$\rho^1_{Ideal} = \rho_1\rho_2\rho_3...\rho_m \ \left(\rho_i = FRR_i \text{ for } i^{th} \text{ sample}\right) \tag{5.13}$$

Here ideal refers to the case of statistically independent decisions. Superscript '0' refers to FAR and '1' to FRR whereas subscripts identify the multiple samples in (5.12) & (5.13). δ 's and ρ 's are the base classifier (1-FAR) and FRR. The equations to calculate the error rates for multi-sample fusion *with incorporation of correlation* between the decisions is given [196] as

$$p^a_{Est} = p^a_{Ideal}\left(1 + \sum_{i<j}\gamma^a_{ij}z^a_i z^a_j + \sum_{i<j<k}\gamma^a_{ijk}z^a_i z^a_j z^a_k + ...\right) \tag{5.14}$$

$$\text{Where } \gamma^a_{123...n} = \sum\left[z^a_1 z^a_2 ....z^a_n\right] \text{ and } z^0_i = \frac{d_i - \delta_i}{\sqrt{\delta_i\left(1-\delta_i\right)}},\ z^1_i = \frac{d_i - \rho_i}{\sqrt{\rho_i\left(1-\rho_i\right)}} \tag{5.15}$$

Here, $\gamma^a (a = 0,1)$ are the correlation coefficients for true speaker and impostor decisions. They are defined using z_i variables that are orthogonal with respect to the independence model with zero mean and unit variance. Decisions d_i are 1 for client & 0 for impostors and so z^1_i are positive for correct client decisions and negative for incorrect ones. For two samples, the first sample results in incorrect decision when the subsequent sample decision is correct most of the time, these variables contribute to negative correlation. Impostor decisions are similarly handled with z^0_i. The magnitude and sign of the correlation, however, depends on the summation over all combinations. The expansion continues to third and higher order decision correlations between classifiers.

For correlated decisions, the ideal error rates calculated using (5.12 & 5.13) under independence assumption are different from experimental error rates or those predicted after applying correlation values in (5.14). The limitation of the analysis presented for sequential decision fusion (sections 2.7 & 4.8) is that the decisions from multiple samples are assumed

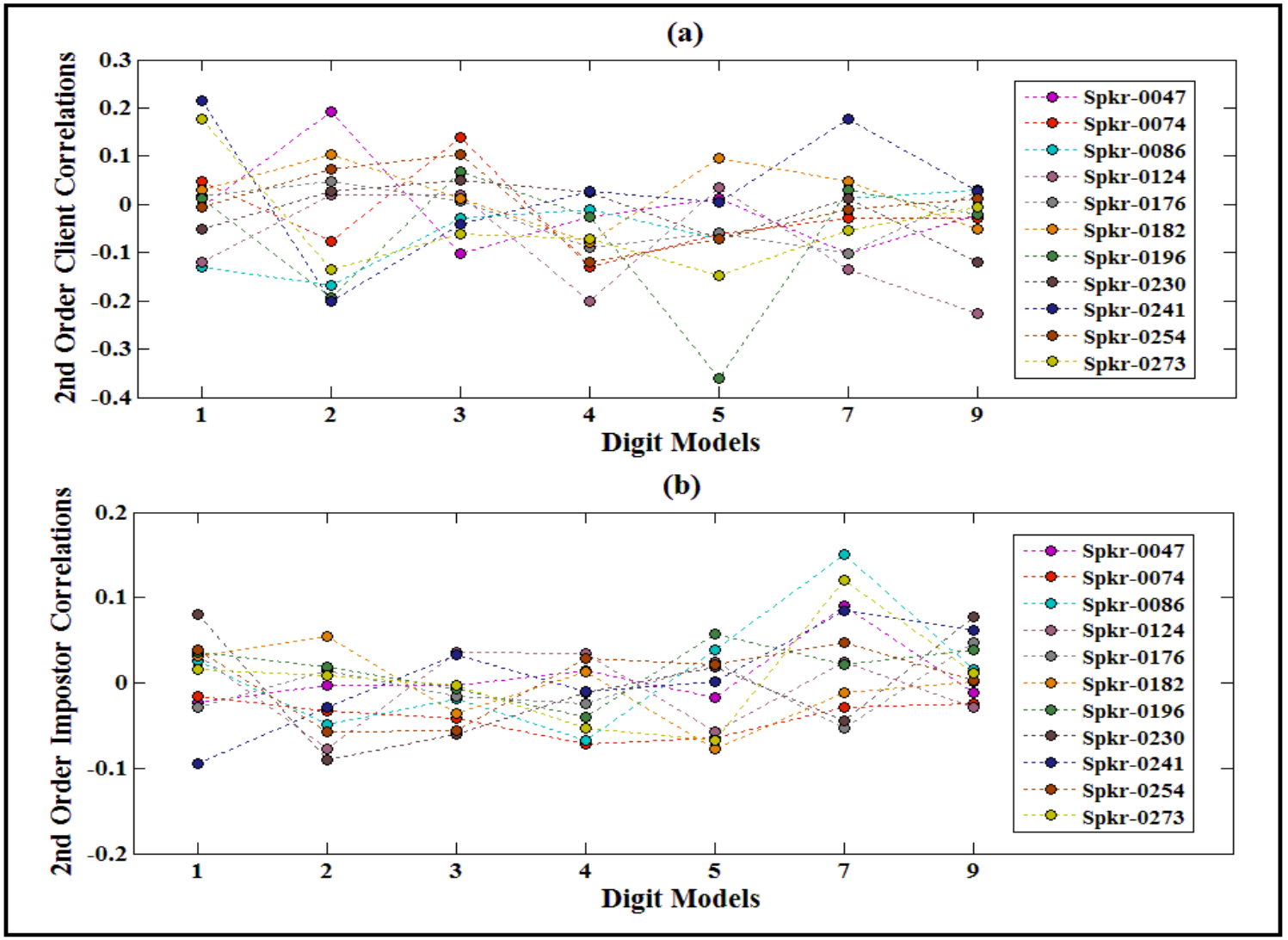

Figure 5.2 2nd-Order Correlations for multi-sample fusion of individual digit models (a) Client and (b) Impostor Decisions

to be independent but this may not be true for multiple samples of an instance/digit spoken by the same individual.

The error rates for the multi-sample fusion calculated under dependence assumption are higher/lower than the experimental values (section 4.8). One of the expected reasons for this difference in error rates is the correlation between decisions from repeated samples. To evaluate the effects of dependence between decisions, the correlation coefficients are calculated using (5.2). The protocol used for the performance evaluation of multi-sample fusion is described in section 3.5.2. A ***sample***, here, refers to an ***utterance*** from a speaker. The ***multi-sample fusion*** is referred as the ***combination of samples***. The scheme in this section is evaluated for speech data from *SET-1*. In this work, the analysis is presented using the pooled results for all the test datasets.

Figure 5.2 presents the 2nd-order correlation between two decisions (first sample and its repetition) of client and impostor random samples from *SET-1*. The 2nd-order correlation coefficients between samples of the same digit for multiple speakers differ in both the sign and magnitude, which may be because of the difference in base performances. The coefficient values for most of the speakers in *SET-1* are negative for client decisions (fig. 5.2(a)) and positive for impostor decisions (fig. 5.2(b)). As the error rates for verification of random sample are considered equal for each repetition, the correlation values estimated

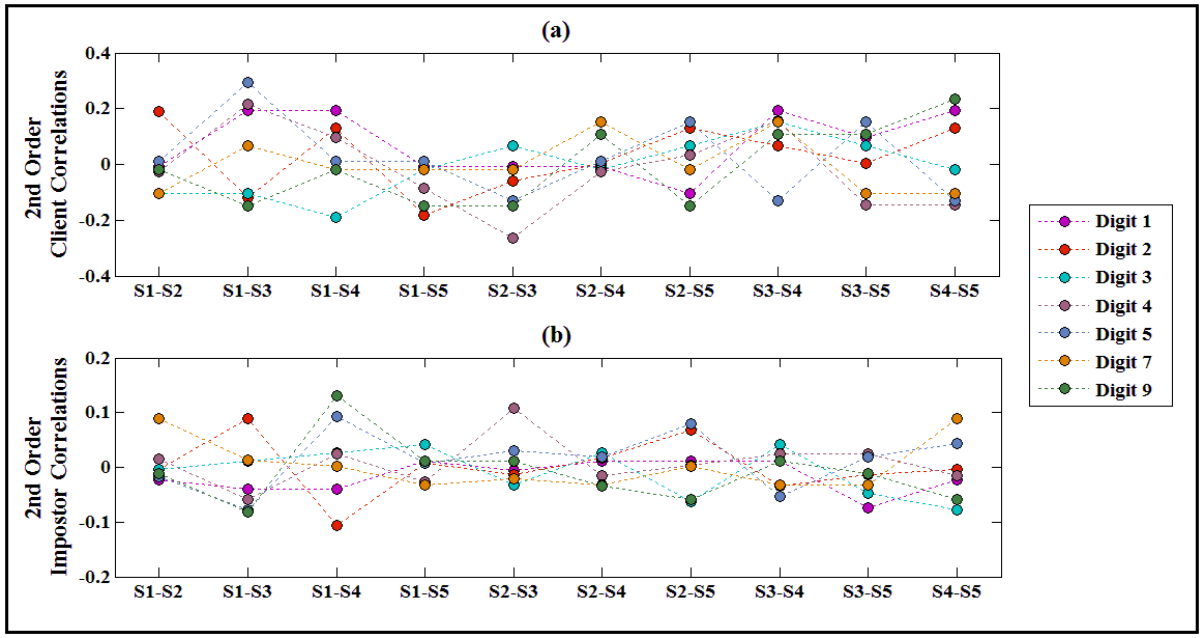

Figure 5.3 2nd-order Correlation for Multi-Sample Fusion of digit models for Speaker-0047 (a) Client Decisions and (b) Impostor Decisions

using (5.4) are supposed to be similar provided the combination errors are equal.

Figure 5.2 presents the 2nd-order correlations for combination of first sample and its randomly repeated sample. For determining the nth-order error rates, it is required to find 2nd-(*n*-1)th order correlation coefficients the samples. Figure 5.3 shows the six possible 2nd-order coefficients for four samples of an instance (digit), i.e., S1-S2 (correlation between Sample 1 and Sample 2) to S3-S4 (correlation between Sample 3 and Sample 4). The correlation values for client and impostors are either similar or different between multiple samples of a digit, i.e. in fig. 5.3 (a), the correlation values for the digit models such as '5' and '9' have similar correlation values whereas for digit models such as '4' and '2' the correlation values are different.

In fig. 5.2 & 5.3, the 2nd-order correlation coefficients between samples of a speaker are represented. The mean correlation values for the 2nd-5th order coefficients are shown in the table 5.2. The ideal error rates (5.12 & 5.13) and correlation values (5.4-5.7) in the table are used to determine the predicted error rates (which when aptly calculated are equal to experimental error rates) for each digit of a speaker. Table 5.2 presents the mean values for the ideal and experimental error rates for the combination of '*m*' samples (*m* = 1, 2, 3, 4, & 5). The *ideal false rejection rates* (FRR) are observed to be slightly lower than the *experimental/predicted FRRs* whereas the *ideal false acceptance rates* (FAR) are somewhat greater than the *experimental/predicted FARs* for multi-sample fusion. This difference in error rates is due to the presence of statistical dependence (positive or negative correlation -

Table 5.2 Ideal and Experimental Error Rates with 2nd-5th order correlations of multi-sample fusion schemes for *SET-1*

Number of Samples	Ideal Error Rates		Experimental Error Rates		Correlation Coefficients	
	FRR	FAR	FRR	FAR	FRR	FAR
2	$0.077^{\pm 0.07}$	$0.399^{\pm 0.22}$	$0.074^{\pm 0.07}$	$0.399^{\pm 0.22}$	$-0.016^{\pm 0.10}$	$-0.001^{\pm 0.05}$
3	$0.029^{\pm 0.02}$	$0.511^{\pm 0.25}$	$0.027^{\pm 0.02}$	$0.511^{\pm 0.25}$	$-0.009^{\pm 0.10}$	$0.003^{\pm 0.05}$
4	$0.012^{\pm 0.01}$	$0.591^{\pm 0.27}$	$0.011^{\pm 0.01}$	$0.592^{\pm 0.27}$	$-0.010^{\pm 0.11}$	$-0.001^{\pm 0.04}$
5	$0.005^{\pm 0.005}$	$0.651^{\pm 0.27}$	$0.005^{\pm 0.005}$	$0.651^{\pm 0.27}$	$-0.006^{\pm 0.10}$	$0.001^{\pm 0.05}$

table 5.3) between the decisions. The correlation coefficients for *client decisions* are demonstrated to be mostly *positive* whereas the *even order correlation coefficients* are *negative* and the *odd order correlation coefficients* are *positive* for impostor decisions. The correlation values for client and impostor decisions are observed to be extremely small and thus the ideal and experimental error rates for multi-sample fusion of random samples are similar. Although the correlation coefficients for random samples are observed to be small, the decisions from adaptive repeated samples can be more dependent. The next section presents the analysis for the dependence between the adaptive samples.

5.2.2.1 Adaptive vs. Random samples

The use of multiple random or adaptive samples reduces the number of false rejects and increases the false accepts. If the sample is rejected, the next sample is either randomly selected or adapted from a rejected sample. In the fusion architecture, if the speaker is accepted by 'i^{th} sample' then the subsequent samples (i+1, i+2 ... m) need not be verified and the fusion performance is thus independent of decisions from these subsequent samples. But for the theoretical estimation of error rates requires the base error rates for each sample used for verification. These base performances between repeated samples are different for adaptive samples whereas equal for random nature of samples. Therefore, the correlation coefficients estimated using these base performances (for random and adaptive samples) are different and so the error differences between the error rates can also be varied.

The statistical difference between the ideal and experimental error rates for multi-sample fusion of random samples is extremely small for samples from *SET-2* (section 4.5). The statistical analysis of *error difference* for random and adaptive is evaluated in this section on three speakers (Spkr-0074, Spkr-0074 & Spkr-0241) from *SET-1*. A test for significance - paired t-test - compares the ideal and experimental error rates for different digit combinations. The results for the paired t-test that compares the error rates for fusion of multiple samples are given in table 5.3. The significance for false rejection and false acceptance rates of fusion of random samples fusion is higher than 0.05 and so the null hypothesis that the difference between the ideal and experimental error rates is zero can be strongly accepted. Since the acceptance is so strong for most of the cases, it can be concluded that all combinations have the same means for both error rates. However, the null hypothesis can be rejected in the case of adaptive samples fusion as the ideal and experimental error rates are significantly different.

Table 5.3 Paired t-test results for Ideal and Experimental Error Rates of multi-sample fusion for adaptive and random samples

		Mean	$S.D^1$	95% Confidence Interval - Difference		t^2	df^3	p^4
				Lower	Upper			
Adaptive Samples	FRR - 2S	-0.024	0.034	-0.026	-0.022	-23.03	1007	<0.0001
	3S	-0.016	0.029	-0.019	-0.013	-10.18	335	<0.0001
	FAR - 2S	0.058	0.041	0.055	0.060	45.12	1007	<0.0001
	3S	0.146	0.176	0.128	0.165	15.24	335	<0.0001
Random Samples	FRR - 2S	0.001	0.020	0.000	0.002	1.63	1007	0.103
	3S	0.001	0.015	-0.001	0.003	1.34	335	0.179
	FAR - 2S	0.000	0.008	-0.001	0.000	-0.87	1007	0.386
	3S	-0.001	0.010	-0.002	0.000	-1.19	335	0.234

1: Standard deviation 2: Paired sample *t*-test

3: Degrees of freedom 4: *p* value. Significance criterion set at $p \leq 0.05$

The empirical evaluation of multi-sample fusion does not consider decisions from the subsequent sample of an accepted sample and thus has no influence on fusion performance. Nevertheless, the correlation coefficients for acceptance or rejection of the subsequent samples are different. To investigate the error differences between false rejection and false acceptances, the correlation values are calculated assuming the subsequent decisions to be 'one' (accept) and 'zero' (reject) for adaptive samples. These values are compared to actual decisions ('Zero/One') for the subsequent adaptive samples. The 2nd and 3rd-order coefficients (5.11) for 'Zero', 'One' and 'Zero/One' subsequent decisions are shown in the table 5.4. The ideal error rates (5.12 & 5.13) and experimental error rates for fusion of two and three samples are presented in this table. The mean ideal error rates are higher/lower than mean experimental error rates. For example, when the subsequent decisions ('Zero/One') are from an actual adaptive sample, the mean false rejection rates (FRR) are lower than mean experimental FRRs and the mean ideal false acceptance rates (FAR) are higher than

Table 5.4 Ideal and Experimental Error Rates with 2nd & 3rd-order correlations for fusion of adaptive samples with the decisions of 'Zero/One', 'Zero' and 'One' for subsequent samples

Adaptive Samples (Subsequent Sample)		Ideal Errors		Experimental Errors		Correlation coefficient	
		FRR	FAR	FRR	FAR	Client	Impostor
'Zero/One'	2S	$0.026^{\pm 0.02}$	$0.457^{\pm 0.27}$	$0.061^{\pm 0.06}$	$0.393^{\pm 0.23}$	$0.295^{\pm 0.23}$	$0.371^{\pm 0.17}$
	3S	$0.003^{\pm 0.003}$	$0.590^{\pm 0.3}$	$0.019^{\pm 0.01}$	$0.492^{\pm 0.27}$	$0.111^{\pm 0.34}$	$-0.225^{\pm 0.32}$
'Zero'	2S	$0.173^{\pm 0.11}$	$0.348^{\pm 0.19}$	$0.061^{\pm 0.06}$	$0.393^{\pm 0.23}$	$0.451^{\pm 0.13}$	$0.629^{\pm 0.14}$
	3S	$0.047^{\pm 0.04}$	$0.484^{\pm 0.25}$	$0.019^{\pm 0.01}$	$0.492^{\pm 0.27}$	$0.694^{\pm 0.36}$	$-0.895^{\pm 0.10}$
'One'	2S	$0.023^{\pm 0.02}$	$0.497^{\pm 0.27}$	$0.061^{\pm 0.06}$	$0.393^{\pm 0.23}$	$-0.861^{\pm 0.14}$	$-0.255^{\pm 0.18}$
	3S	$0.001^{\pm 0.001}$	$0.672^{\pm 0.29}$	$0.019^{\pm 0.01}$	$0.492^{\pm 0.27}$	$-0.463^{\pm 0.38}$	$0.019^{\pm 0.14}$

experimental FARs for multi-sample fusion. The *error difference* here is because of higher statistical dependence (correlation value in table 5.4) between adaptive sample decisions. For 'Zero/One' case, the adaptive decisions for clients are positively dependent whereas for impostor adaptive decisions the 2nd-order coefficients are positive and 3rd-order coefficients are negative. For the three cases in table 5.4, it is demonstrated that correlation values are either higher or lower than 'zero' thereby indicating that the ideal and experimental error rates in these case are also not equal. Thereby, the error difference is used to distinguish between adaptive (error difference > or < 'Zero') or random (error difference $\approx$ 'Zero') nature of the repeated sample. Based on these correlation values and base error rates, the decisions for repeated samples of a digit with favourable/unfavourable dependence can be determined.

5.2.3 Fusion of '*n*' instances and '*m*' samples

The architecture for integration of multi-instance and multi-sample fusion schemes is presented in the section 4.5. The architecture is explained using the logical rules. For a speaker to be declared genuine for a particular instance, it is considered sufficient if any one sample (or utterance) presented to the system gets accepted. Acceptance decisions are logical *'OR'* for multiple samples. The speaker is considered to be an impostor when all the '*m*' samples are rejected. Rejection decisions are logical *'AND'* for multiple samples. Conversely, it is considered necessary in the sequential decision framework that a speaker be accepted by

all instances in the sequence of decision stages. Acceptance is thus logical *'AND'* for multiple instances. If the speaker is rejected by any decision stage, the sequence terminates and thus rejection decisions are logical *'OR'* for multiple instances.

5.2.3.1 False Accepts

The decisions from multiple samples $d_{Si}(i=1,2,3,...m)$ are combined in the fusion process to obtain a decision about the identity claim of the speaker at an instance level. Such decisions from multiple instances $d_{Sm}^{Ci}(i=1,2,3,...n)$ are combined to obtain a final accurate decision about the identity claim of the speaker. The false acceptance rate, $\alpha_{S1,S2,...Sm}^{Cn}$, for the fusion of '*m*' samples (*S1, S2,... Sm*) of an '*n*th' instance is given as

$$\alpha_{S1,S2,..Sm}^{Cn}=\alpha_{Ideal}^{Cn}\left(1+\sum_{i<j}\gamma_{ij}^{0}\sqrt{\frac{\alpha_{S1}^{Cn}\alpha_{S2}^{Cn}}{\left(1-\alpha_{S1}^{Cn}\right)\left(1-\alpha_{S2}^{Cn}\right)}}+\sum_{i<j<k}\gamma_{ijk}^{0}\sqrt{\frac{\alpha_{S1}^{Cn}\alpha_{S2}^{Cn}\alpha_{S3}^{Cn}}{\left(1-\alpha_{S1}^{Cn}\right)\left(1-\alpha_{S2}^{Cn}\right)\left(1-\alpha_{S3}^{Cn}\right)}}+....+\gamma_{123..n}^{0}\right)$$

Where $\alpha_{Ideal}^{Cn}=\alpha_{S1}^{Cn}+\left(1-\alpha_{S1}^{Cn}\right)\alpha_{S2}^{Cn}+....+\left(1-\alpha_{S1}^{Cn}\right)\left(1-\alpha_{S2}^{Cn}\right)....\left(1-\alpha_{S(m-1)}^{Cn}\right)\alpha_{Sm}^{Cn}$

The claim is declared genuine, at the end of '*n*' instances, if accepted at all the instances (*'AND rule'*). The FAR for fusion of decisions from multiple instances, *i*=1, 2, 3 ...n, $(\alpha_{S1,S2,..Sm}^{C1,C2,..Cn})$ using '*AND'* logic is given as

$$\alpha_{S1,S2,...Sm}^{C1,C2,...Cn}=\alpha_{I}\left(1+\sum_{i<j}\gamma_{ij}^{1}\sqrt{\frac{\left(1-\alpha_{S1,S2,...Sm}^{Ci}\right)\left(1-\alpha_{S1,S2,...Sm}^{Cj}\right)}{\alpha_{S1,S2,...Sm}^{Ci}\alpha_{S1,S2,...Sm}^{Cj}}}+\sum_{i<j<k}\gamma_{ijk}^{1}\sqrt{\frac{\left(1-\alpha_{S1,S2,...Sm}^{Ci}\right)\left(1-\alpha_{S1,S2,...Sm}^{Cj}\right)\left(1-\alpha_{S1,S2,...Sm}^{Ck}\right)}{\alpha_{S1,S2,...Sm}^{Ci}\alpha_{S1,S2,...Sm}^{Cj}\alpha_{S1,S2,...Sm}^{Ck}}}+...\right) \quad (5.16)$$

where $\alpha_{I}=\alpha_{S1,S2,...Sm}^{C1}\ \alpha_{S1,S2,...Sm}^{C2}\ \alpha_{S1,S2,...Sm}^{C3}.....\ \alpha_{S1,S2,...Sm}^{Cn}$

5.2.3.2 False Rejects

If the speaker is rejected for all the repeated samples of an instance, the client's claim is rejected. The false rejects for fusion of '*m*' multiple samples is determined using the *'AND'* logic and is expressed as

$$\rho^{Cn}_{S1,S2,S3.....,Sm} = \rho^{Cn}_{S1}\rho^{Cn}_{S2}\rho^{Cn}_{S3}\cdots\rho^{Cn}_{Sm}\left(1+\sum_{S1<S2}\gamma^{0}_{S1S2}\sqrt{\frac{\left(1-\rho^{C1}_{S1}\right)\left(1-\rho^{C1}_{S2}\right)}{\rho^{Cn}_{S1}\ \rho^{Cn}_{S2}}}+\sum_{S1<S2<S3}\gamma^{0}_{S1S2S3}\sqrt{\frac{\left(1-\rho^{C1}_{S1}\right)\left(1-\rho^{C1}_{S2}\right)\left(1-\rho^{C1}_{S3}\right)}{\rho^{Cn}_{S1}\ \rho^{Cn}_{S2}\ \rho^{Cn}_{S3}}}+....\right)$$

If the speaker is rejected at any instance, the client claim is rejected. The false rejects for fusion of '*n*' multiple instances, determined using the *'OR'* logic expressions, are obtained by substituting error rates in the equation for multi-instance fusion with that of expressions for multi-sample fusion.

$$\rho^{C1,C2,...Cn}_{S1,S2,...Sm} = \rho_I\left(1+\sum_{i<j}\gamma^{0}_{ij}\sqrt{\frac{\rho^{C1}_{S1,S2,...Sm}\rho^{C2}_{S1,S2,...Sm}}{\left(1-\rho^{C1}_{S1,S2,...Sm}\right)\left(1-\rho^{C2}_{S1,S2,...Sm}\right)}}+\sum_{i<j<k}\gamma^{0}_{ijk}\sqrt{\frac{\rho^{C1}_{S1,S2,...Sm}\rho^{C2}_{S1,S2,...Sm}\rho^{C3}_{S1,S2,...Sm}}{\left(1-\rho^{C1}_{S1,S2,...Sm}\right)\left(1-\rho^{C2}_{S1,S2,...Sm}\right)\left(1-\rho^{C3}_{S1,S2,...Sm}\right)}}+...\right) \quad (5.17)$$

where $\rho_I = \rho^{C1}_{S1,S2,..Sm}+\left(1-\rho^{C1}_{S1,S2,..Sm}\right)\rho^{C2}_{S1,S2,..Sm}+.....$
$$+\left(1-\rho^{C1}_{S1,S2,..Sm}\right)\left(1-\rho^{C2}_{S1,S2,..Sm}\right)...\left(1-\rho^{C(n-1)}_{S1,S2,..Sm}\right)\rho^{Cn}_{S1,S2,..Sm}$$

The error rates for fusion of multiple instances and multiple samples calculated under dependence assumption are higher/lower than experimental values. One of the expected reasons for this difference in error rates is the correlation between classifier decisions. To evaluate the influence of dependence between classifier decisions, the correlation coefficients are calculated using (5.11). The correlation coefficients for fusion of decisions from multiple instances and multiple samples are shown in table 5.5. The experimental error rates and ideal error rates (calculated using base performances) used for calculation of these coefficients (5.11) is presented in the table. The ideal false rejection rates (FRR) are higher than the experimental FRRs whereas the ideal false acceptance rates (FAR) are lower than the experimental FARs for proposed fusion. This *error difference* is due to the presence of statistical dependence (correlation value in table 5.5) between client and impostor decisions. The error difference decreases with an increase in digits and samples used for fusion as the correlation progressively decreases to zero. The increase in repeated samples can result in

Table 5.5 Ideal and Experimental Error Rates for Sequential Decision Fusion Scheme with Correlation Coefficients

	Ideal Error Rates		Experimental Error Rates		Correlation Coefficients	
	FRR	FAR	FRR	FAR	Client	Impostor
1D-1S	$0.237^{\pm 0.14}$	$0.238^{\pm 0.14}$	$0.237^{\pm 0.14}$	$0.238^{\pm 0.14}$	-	-
2D-2S	$0.140^{\pm 0.12}$	$0.190^{\pm 0.16}$	$0.136^{\pm 0.12}$	$0.200^{\pm 0.16}$	$0.051^{\pm 0.15}$	$0.052^{\pm 0.07}$
3D-3S	$0.077^{\pm 0.07}$	$0.193^{\pm 0.18}$	$0.076^{\pm 0.07}$	$0.203^{\pm 0.18}$	$0.001^{\pm 0.18}$	$0.005^{\pm 0.05}$
4D-4S	$0.044^{\pm 0.04}$	$0.211^{\pm 0.21}$	$0.043^{\pm 0.04}$	$0.219^{\pm 0.21}$	$0.001^{\pm 0.04}$	$0.001^{\pm 0.05}$
5D-5S	$0.023^{\pm 0.02}$	$0.235^{\pm 0.23}$	$0.023^{\pm 0.02}$	$0.241^{\pm 0.24}$	$-0.001^{\pm 0.001}$	$-0.001^{\pm 0.05}$

zero false rejects when the base error rates are small or the number of tests performed may not be sufficient to represent errors at required precision. The use of these errors in correlation estimation can result in undefined values. The mean correlation values represented in table 5.5 excludes the digit combinations with undefined values. Though the differences between ideal and experimental values are lowered with an increase in instances and samples used for fusion, significant improvement in fusion performance is obtained using digit combinations with favourable dependence. The next section presents analysis to determine the solution for favourable dependence of '*n*' instance and '*m*' sample fusion.

5.3 Analysis of favourable/unfavourable dependence between decisions

OVERVIEW: This subsection analyses the 'favourable/unfavourable' dependence for 'n' instance or 'm' sample fusion, i.e., multi-instance and multi-sample decision fusion schemes. Dependence for the general case of 'k' classifier fusion is difficult because of the coupling between dependence of kth classifier dependence and dependence of all 1, 2... (k − 1)th classifier decisions. The previous work on decision fusion using 'AND' fusion has shown that the dependence is favourable if even-order correlation coefficients should be negative and

odd-order correlation coefficients should be positive for impostor decisions. Whereas 2nd-order negative client correlation coefficient and a 2nd-order positive impostor correlation coefficient is favourable for sequential 'OR' fusion of decisions from two samples. However, this analysis was based on only the signs of the correlation coefficients and was not a complete solution.

This section provides the theoretical analysis for the general case of 'n' classifier decision fusion. It is shown that the dependence between the decisions was determined based on an error factor that includes the base errors and magnitude of correlation between decisions when the coefficients are of different signs. These developed expressions for favourable dependence are experimentally evaluated using the sequential architecture for text-dependent speaker verification. The multi-instance fusion performance is demonstrated to be better when impostor and client-impostor favourable digit combinations are considered because FAR decreases with increase in digits. The multi-sample fusion performance is demonstrated to be better when client and client-impostor favourable digit combinations are considered because FRR decreases with increase in digits. The client-impostor favourable combinations also ensure that the experimental or predicted error rates (fusion of favourable dependent decisions) are always lower than the ideal error rates (fusion of independent decisions).

Tumer and Ghosh [73] considered the use of correlation coefficient as the diversity measure to reduce the error. For negative values of correlation coefficients, the added error of the mean (sum) score combiner is shown to be smaller than the added error of statistically independent classifiers. The results obtained were observed to be misleading, as the analysis does not deal with the class-conditional errors and class-conditional diversity values. Venkataramani [39] investigated the questions related to diversity using joint probability distributions that help in representing the optimal decision fusion accuracy as a function of statistical dependence and design of classifier ensembles (using favourable joint probability distributions). This analysis also considers the class-conditional errors and class-conditional diversity values for *AND*, *OR* and majority voting decision fusion rules. The analysis in this dissertation follows the work in [39] and investigates the relation between the class conditional error rates of the proposed fusion method and class-conditional dependence values. The relation between error rates and dependence can be summarized as:

- For classifier decisions with *statistical independence*, the error rates for *'AND'* and 'OR' rules are calculated using base classifier error rates. When the classifier decisions are *statistically dependent*, the error rates after decision fusion may be larger or smaller than when the classifier decisions are *statistically independent*.
- If the error rates after fusion using either *'AND'* or *'OR'* rules were smaller than those of independent classifier decisions, then the dependence is termed as "*favourable*"
- If the error rates for dependent decision fusion were higher than that of independent classifier decisions, then the dependence is '*unfavourable*'
- The dependence is considered '*optimal*' for a given fusion rule, when the error rates for a particular combination of decisions is the lowest compared to other decision combinations.

The analysis in [39], however, does not provide the complete set of conditions to identify the favourable or optimal dependence for '*n'* decisions. The subsequent sections deal with the correlation analysis for multi-instance and multi-sample fusion schemes that employ the *'AND'* and *'OR'* fusion rules.

5.3.1 Multi-instance Fusion

In *'Evaluation and Selection'* method of verification, the error rates for the test dataset are predicted using the parameters tuned on the development/tune dataset. When correlated decisions are fused, the parameters to be tuned include the correlation values, which, in general, are data dependent. As the test dataset is assumed unknown, the actual correlation values are not known. Therefore, the error rates for the test dataset can be estimated using the independence assumption, provided the parameters selected on the development dataset ensure that the dependence (positive or negative) between the decisions result in comparably lower error rates than independence assumption. This method of parameter selection ensures that the predicted error rates (for correlated decisions) are lower than the ideal error rates (for independent decisions). The dependence condition for which the error rates after fusion of correlated decision using fusion rules (e.g., *'AND*' or *'OR'*) are smaller than those of independent classifier decisions is termed as "*favourable*". The analysis of *favourable dependence* is also used to determine the better relationship between the *correlation coefficients* and the *Error Difference*.

The relationship between favourable dependence and correlation between the classifier decisions depends on the fusion rule. Venkataramani and Vijaya Kumar [37] has

shown that for *'AND rule'* the positive correlation coefficients are favourable for client decisions whereas and negative correlation values are favourable for impostor decisions. The favourable dependence here implies that the error difference is greater than zero, i.e., the difference between ideal error rates and experimental error rates are greater than zero. The analysis in [37] is empirically evaluated for the dependence between the decisions from multiple instances for text-dependent speaker verification. Figures 5.4 (a) & (b) presents the 2nd-order client and impostor coefficients for combination of two digits from three speakers (Spkr-0074, Spkr-0074 & Spkr-0241) in *SET-1*. The negative error difference in the figure represents that ideal error rates are lower than experimental error rates. The 2nd-order coefficients are shown to have strong positive dependence with an error difference in FRR, i.e., with the increase in correlation the FRR error difference increases. Moreover, the coefficients have strong negative dependence with error difference in FAR i.e., the decrease in correlation increases the error difference in FAR.

From the expression of correlation coefficients for *'n'* decisions (5.11), it is evident that the nth order correlation coefficient is dependent on 2^{nd}, 3^{rd} . . . , $(n-1)$th order coefficients (for example, (5.5 & 5.6) represent the expressions for 3rd and 4th-order coefficients represented in terms of lower order coefficients and base performances). The dependence on lower order correlation coefficients results in a weak relationship between

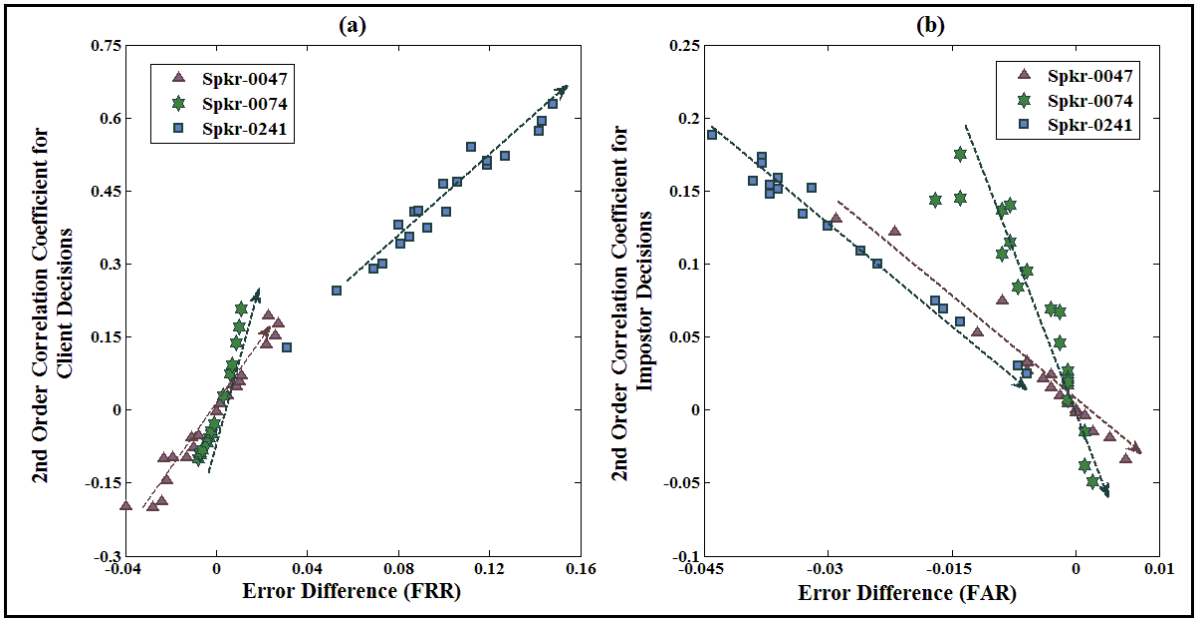

Figure 5.4 Error Differences and 2nd-Order Correlation Coefficients for Multi-Instance Fusion of (a) Client Decisions and (b) Impostor Decisions

error difference and nth order coefficients. Figure 5.5 (a) and (c) shows this lack of relationship between error difference and correlation coefficients (client and impostor decisions) for three-digit combinations of Spkr-0074, Spkr-0074 and Spkr-0241 from *SET-1*. The dependence for *'n'* decisions (n>2) is determined by relaxing the relationship of *n*th order coefficients with a lower order (2, 3, 4 ... (n-1)) correlation coefficients. Figure 5.5(b) and (d) plots the 3rd-order correlation coefficients and error difference for client and impostor decisions of the same speakers in fig. 5.4 (a) and (c) respectively. The 3rd-order coefficients are calculated here under the assumption of *'zero'* 2nd-order correlation coefficients for client and impostor decisions. With this assumption, the positive 3rd-order client correlation coefficients and negative 3rd-order impostor correlation coefficients are shown to be favourable for *'AND'* fusion rule (fig 5.5(b) and (d)). The same conclusion could be extended

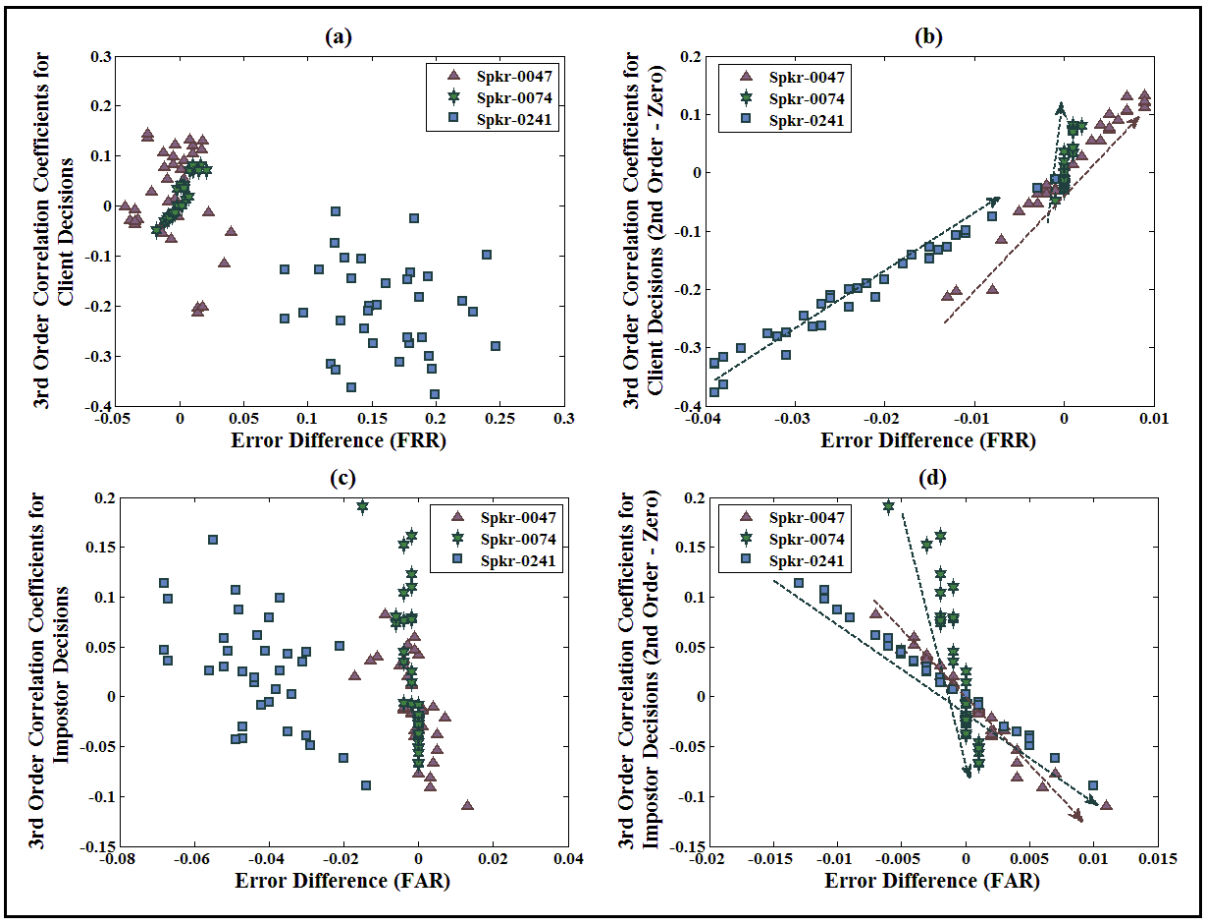

Figure 5.5 Error Differences and 3rd-order Correlation Coefficients for Multi-Instance Fusion of (a) Three client decisions (b) Three client decisions with 'Zero' 2nd-order coefficients between two client decisions (c) Three impostor decisions and (d) Three impostor decisions with 'Zero' 2nd-order coefficients between two client decisions

to *n*th order coefficients (*n*>3) where the favourable dependence is determined under the assumption that the correlation coefficients between (*n*-1)... 2 client and impostor decisions are considered to be zero.

The expressions for the error rates (5.12 & 5.13) are composed of all the *k*th-order coefficients (5.5-5.7 & 5.11), $k = 2, 3. . . n$ and so the relative weight of these terms play a role in the determining the favourable dependence for *AND rule*. Venkataramani [39] presented the analysis to identify the favourable dependence for *n*th order coefficients when the decisions from 2, 3... (*n*-1) classifier combinations are not statistically independent and are combined using *'AND or OR'* fusion rules [39]. For '*n*' classifier *AND fusion*, the positive even-order correlation coefficients and negative-odd order correlation coefficients are considered to be favourable on client decisions. The negative even-order correlation coefficients and positive odd-order correlation coefficients are shown to be favourable on impostor decisions. This condition of favourable dependence is entirely based on the signs of correlation coefficients and considered as only one of the sufficient conditions for dependence.

The condition given in [39] is empirically evaluated to investigate the conditions of dependence (favourable/unfavourable) between classifier decisions. Figure 5.6 plots correlation coefficients and error differences for the combination of five digits 2-3-4-7-9 from Spkr-0047 in *SET-1*. The figure represents correlation coefficients for various digit combinations - two digits (23, 24, 27 ... 79), three digits (234, 237 ... 479), four digits (2347, 2349 ... 3479) and five digits (23479) that can be either favourable (positive error difference) or unfavourable (negative error difference). The points to be considered from figure 5.6 (a) and (b) for client and impostor correlated decisions respectively are:

- From [39], for general case of '*n*' classifier *'AND fusion'* the dependence is favourable when even-order coefficients are positive and odd-order correlation coefficients are negative for client decisions. For impostor decisions, even-order and odd-order correlation coefficients should be negative and positive respectively. The analysis is based on determining the conditions for which *k*th-order probabilities of dependent classifiers are better than independent classifier values. These conditions are derived by investigating the sign for each correlation coefficient (e.g., 2nd-order, 3rd-order coefficients) independently. The probability of detection or false alarm of '*n'*th order, however, depends on the combined probabilities of

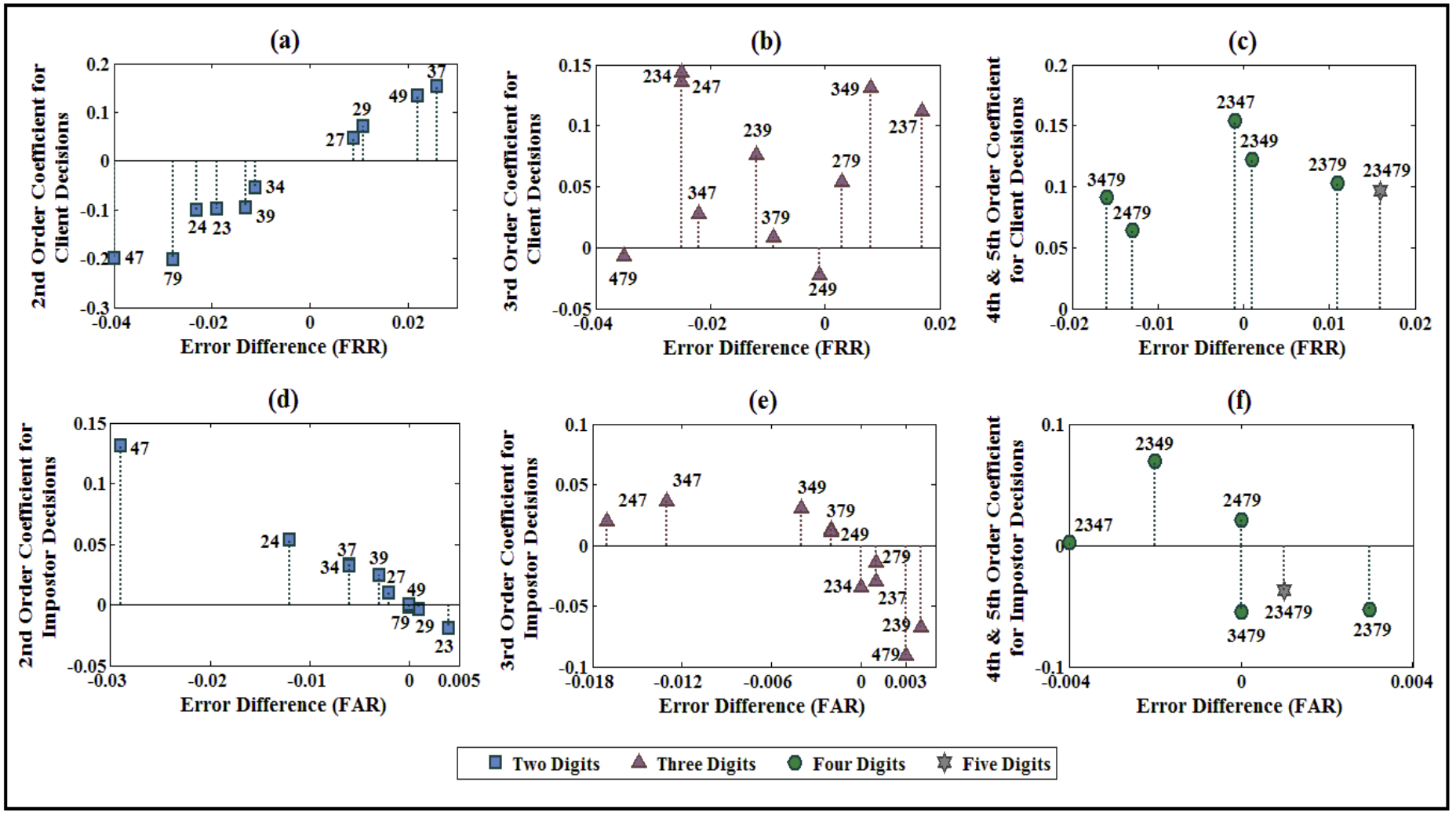

Figure 5.6 Correlation Coefficients and Error Differences for combinations of digits 2-3-4-7-9 for Client and Impostor Decisions of Spkr-0047 from *SET-1*

'*i*'th order (*i* =1, 2, 3 ... (*n*-1)). The analysis does not include the condition for favourable dependence when correlation coefficients of a particular order are not of the same sign. For example, prediction of error rates for the combination of five digits 2-3-4-7-9 depends on the probabilities (and correlation coefficients) of *k*th order (*k*=1, 2, 3 & 4) where all even order (2nd & 4th) coefficients are not positive and odd order (3rd) coefficients are not negative in fig 5.6.(a), (b) & (c).

- It is possible that conditional dependence between the decisions for *'AND rule'* is favourable when some of the even-order correlation coefficients are negative and some of the odd-order correlation coefficients are positive for client decisions. Similarly, positive even-order correlation coefficients and negative odd-order correlation coefficients are favourable for impostor decisions.
- The '*n*'th order correlation coefficient for combination of decisions can be favourable, even when the corresponding lower order correlation coefficients may be unfavourable. For example, the dependence between the impostor decisions from three digits (4-7-9) is observed to be favourable although the correlation for two-digit combinations (47, 49 and 79) is positive and unfavourable (fig 5.6(d)).

Therefore, the sign and the magnitude of correlation coefficients are required to determine absolute solution for the favourable/unfavourable dependence of decisions.

5.4.1.1 Favourable dependence for '*n*' impostor decisions

For statistical favourable dependence condition, the predicted false acceptance rate (FAR) for correlated decisions is lower than errors calculated under an independence assumption for the *AND* fusion rule. The equation to satisfy above condition is given as

$$\alpha_P < \alpha_{Ideal} \tag{5.18}$$

$$\alpha_{Ideal}\left(1+\sum_{i<j}\gamma_{ij}^0 z_i^0 z_j^0+\sum_{i<j<k}\gamma_{ijk}^0 z_i^0 z_j^0 z_k^0+\ldots\right)<\alpha_{Ideal} \quad \text{(from equation 5.10)} \tag{5.19}$$

The equation is also described as the condition where '*Error Difference' in FAR* is greater than zero. In (5.19), the dependence between the impostor decisions is considered favourable when the correlation factor $\left(\sum_{i<j}\gamma_{ij}^0 z_i^0 z_j^0+\sum_{i<j<k}\gamma_{ijk}^0 z_i^0 z_j^0 z_k^0+\ldots\right)$ for the fusion of dependent

decisions is negative. The conditions of 'favourable/unfavourable' statistical dependence for the *AND* fusion of '*n*' instances are determined here with the assumption of equal FAR $\alpha_i = \alpha (i = 1, 2, 3, ..n)$ for n instances. The simplified correlation factor with expansion of normalised variables (5.11) for false accepts (i.e., $d_i = 1$) is given as

$$\left(\sum_{i<j} \gamma_{ij}^0 \left(\frac{1-\alpha}{\alpha} \right)^{\frac{2}{2}} + \sum_{i<j<k} \gamma_{ijk}^0 \left(\frac{1-\alpha}{\alpha} \right)^{\frac{3}{2}} + .. + \gamma_{123..n}^0 \right) < 0 \tag{5.20}$$

When the correlation factor is of positive sign, the dependence between the decisions is unfavourable as the predicted FAR for 'AND' fusion of '*n*' correlated decisions is larger than the fusion of '*n*' independent classifier decisions.

5.4.1.1.1 Two Classifier AND Rule

For the combination of two impostor decisions (*n*=2), the above (5.19) is reduced to

$$\gamma_{12}^0 \left(\frac{1-\alpha}{\alpha} \right) < 0 \tag{5.21}$$

Since $\alpha \leq 1$, the error factor in (5.15) is either zero or positive (undefined value when $\alpha = 0$). The only condition for satisfying above equation is when the 2nd-order correlation coefficient is negative $(\gamma_{12}^0 < 0)$. Therefore, the dependence between two impostor decisions is favourable when the 2nd-order correlation coefficient is negative and unfavourable when the 2nd-order correlation coefficient is positive.

5.4.1.1.2 Three Classifier AND Rule

For the combination of three classifier (*n*=2) decisions from impostors, the equation 5.19 is given as

$$\left(\sum_{i<j} \gamma_{ij}^0 \left(\frac{1-\alpha}{\alpha} \right) + \gamma_{ijk}^0 \left(\frac{1-\alpha}{\alpha} \right)^{\frac{3}{2}} \right) < 0$$

$$\left(\gamma_{12}^0 \left(\frac{1-\alpha}{\alpha} \right) + \gamma_{13}^0 \left(\frac{1-\alpha}{\alpha} \right) + \gamma_{23}^0 \left(\frac{1-\alpha}{\alpha} \right) + \gamma_{123}^0 \left(\frac{1-\alpha}{\alpha} \right)^{\frac{3}{2}} \right) < 0$$

$$\left(\gamma_{12}^{0}+\gamma_{13}^{0}+\gamma_{23}^{0}+\gamma_{123}^{0}\sqrt{\frac{1-\alpha}{\alpha}}\right)<0 \tag{5.22}$$

When the 2nd and 3rd-order correlation coefficients are of the same sign and positive, the correlation factor in (5.22) is positive and the condition for favourable dependence is not satisfied. If the 2nd-order and 3rd-order correlation coefficients are of same sign and negative, the condition in (5.22) is justified and the dependence between decisions is favourable. Thus, the *negative 2nd and 3rd-order coefficients* are favourable whereas *positive 2nd and 3rd-order coefficients* are unfavourable.

When the 2nd and 3rd-order coefficients are of different signs, the condition for determining favourable dependence depends on both the correlation coefficients and base FAR of the instances. For positive 2nd-order coefficients and negative 3rd-order coefficient, the condition for favourable dependence (5.22) is given as

$$\left(\gamma_{123}^{0}\sqrt{\frac{1-\alpha}{\alpha}}\right)<\left|\gamma_{12}^{0}+\gamma_{13}^{0}+\gamma_{23}^{0}\right| \tag{5.23}$$

The condition above is reversed for the case with positive 2nd-order coefficients and negative 3rd-order coefficient.

$$\left(\gamma_{12}^{0}+\gamma_{13}^{0}+\gamma_{23}^{0}\right)<\left|\gamma_{123}^{0}\sqrt{\frac{1-\alpha}{\alpha}}\right| \tag{5.24}$$

For positive 2nd-order coefficients and negative 3rd-order coefficient, the decisions are favourable when sum of 2nd-order coefficients is less than the product of the 3rd-order correlation and the error factor $\left(\gamma_{123}^{0}\sqrt{\frac{(1-\alpha)}{\alpha}}\right)$. Although positive 2nd-order coefficients are unfavourable (5.22) for instance fusion, the 3rd-order coefficient calculated from these 2nd-order coefficients can be favourable for fusion of three instances (e.g., in fig 5.6). When 2nd (even) order coefficients are negative and 3rd (odd) order coefficients are positive, the decisions are supposed to be favourable [39]. Nevertheless, the analysis based on signs may not be reliable for all values of correlation, i.e., when sum of the 2nd-order coefficients is less than the product of the 3rd-order correlation and error factor.

5.4.1.1.3 'n' Classifier AND Rule

The analysis for favourable dependence of decisions from *n* instances is similar to that of three instances. The condition (5.20) for favourable dependence of '*n'* instance decisions is given as

$$\left(\sum_{i<j} \gamma_{ij}^{0} + \sum_{i<j<k} \gamma_{ijk}^{0} \left(\frac{1-\alpha}{\alpha} \right)^{\frac{1}{2}} + \sum_{i<j<k<l} \gamma_{ijkl}^{0} \left(\frac{1-\alpha}{\alpha} \right)^{\frac{2}{2}} + ... + \gamma_{123..n}^{0} \left(\frac{1-\alpha}{\alpha} \right)^{\frac{n-2}{2}} \right) < 0 \qquad (5.25)$$

From the above equation, it is evident that the 2nd-nth order correlation coefficients of the same sign and negative are favourable whereas coefficients with same sign and positive are unfavourable. For the correlation coefficient with different signs, the condition for favourable dependence depends on the 2nd-nth order correlation coefficients and the FAR (α) for the fusion of '*n*' classifier decisions.

The generalized equation for determining the favourable dependence between impostor decisions of '*n*' instances combined using *'AND' rule* is obtained by relaxing the assumption of equal FAR (α). Considering individual FAR $\alpha_i (i = 1, 2, 3, ..n)$ for '*n*' instances (5.19) is expanded as

$$\left(\sum_{i<j} \gamma_{ij}^{0} \prod_{k=1}^{n} \prod_{k\neq i, k\neq j} \left(\sqrt{\frac{\alpha_k}{1-\alpha_k}} \right) + \sum_{i<j<k} \gamma_{ijk}^{0} \prod_{l=1}^{n} \prod_{l\neq i, l\neq j, l\neq k} \left(\sqrt{\frac{\alpha_l}{1-\alpha_l}} \right) + ... + \gamma_{123...n}^{0} \right) < 0 \qquad (5.26)$$

The 2nd-*n*th order correlation coefficients of the same sign and negative are favourable whereas coefficients with same sign and positive are unfavourable. When the coefficients are of different signs, the dependence between the decisions is determined using the base FAR and magnitude of the correlation coefficients, i.e., the sum of (*n*-1)th order correlation coefficients multiplied with corresponding FAR factor $\left(\frac{\alpha_n}{1-\alpha_n} \right)$ of *n*th instance.

5.4.1.2 Favourable dependence for '*n*' client decisions

The favourable dependence for multi-instance fusion of client decisions is analysed in steps similar to fusion of impostor decisions. The generalized equation for favourable dependence of client decisions from '*n*' instances with equal FRR of ' ρ ' is given as

$$\left(\sum_{i<j}\gamma_{ij}^{1}+\sum_{i<j<k}\gamma_{ijk}^{1}\left(\frac{\rho}{1-\rho}\right)^{\frac{1}{2}}+\sum_{i<j<k<l}\gamma_{ijkl}^{1}\left(\frac{\rho}{1-\rho}\right)^{\frac{2}{2}}+.....+\gamma_{123..n}^{1}\left(\frac{\rho}{1-\rho}\right)^{\frac{n-2}{2}}\right)>0 \qquad (5.27)$$

From the above equation, it is evident that the 2nd-nth order correlation coefficients of the same sign and positive are favourable whereas coefficients with same sign and negative are unfavourable. For the correlation coefficient with different signs, the condition for favourable dependence depends on the 2nd-nth order correlation coefficients and the FRR (ρ) for the fusion of '*n*' classifier client decisions.

The generalized equation for determining the favourable dependence between impostor decisions of '*n*' instances combined using *'AND' rule* is obtained by relaxing the assumption of equal FRR (ρ). Considering individual FRR $\rho_i (i = 1, 2, 3, 4...n)$ for '*n*' instances the (5.27) is expanded as

$$\left(\sum_{i<j}\gamma_{ij}^{1}\prod_{k=1}^{n}\prod_{k\neq i,k\neq j}\left(\sqrt{\frac{1-\rho_k}{\rho_k}}\right)+\sum_{i<j<k}\gamma_{ijk}^{1}\prod_{l=1}^{n}\prod_{l\neq i,l\neq j,l\neq k}\left(\sqrt{\frac{1-\rho_l}{\rho_l}}\right)+....+\gamma_{123...n}^{1}\right)>0 \qquad (5.28)$$

From the above equation, it is evident that the 2nd-nth order correlation coefficients of the same sign are favourable when positive and unfavourable when negative. If the coefficients are of different signs, the dependence between the decisions is determined using the base FRR and magnitude of correlation between client decisions, i.e., the sum of (*n*-1)th order correlation coefficients multiplied with FRR factor $\left(\frac{1-\rho_n}{\rho_n}\right)$ of the *n*th instance. The ***favourable dependence*** between the decisions from either client or impostors enables to determine the ***best set of classifier combinations*** that result in lower error rates for fusion of dependent decisions rather than fusion of independent decisions using sequential *'AND rule'*.

5.3.1.3 Error rates for favourable digit combinations

The analysis on favourable dependence enables to determine the best set of classifiers with *Error Difference* greater than zero (i.e., fusion of dependent decisions results in lower error rates than fusion of independent decisions). The client favourable and impostor favourable combinations for text-dependent speaker verification are evaluated using (5.28) and (5.26) respectively. The protocol used for evaluation of multi-instance fusion (digit

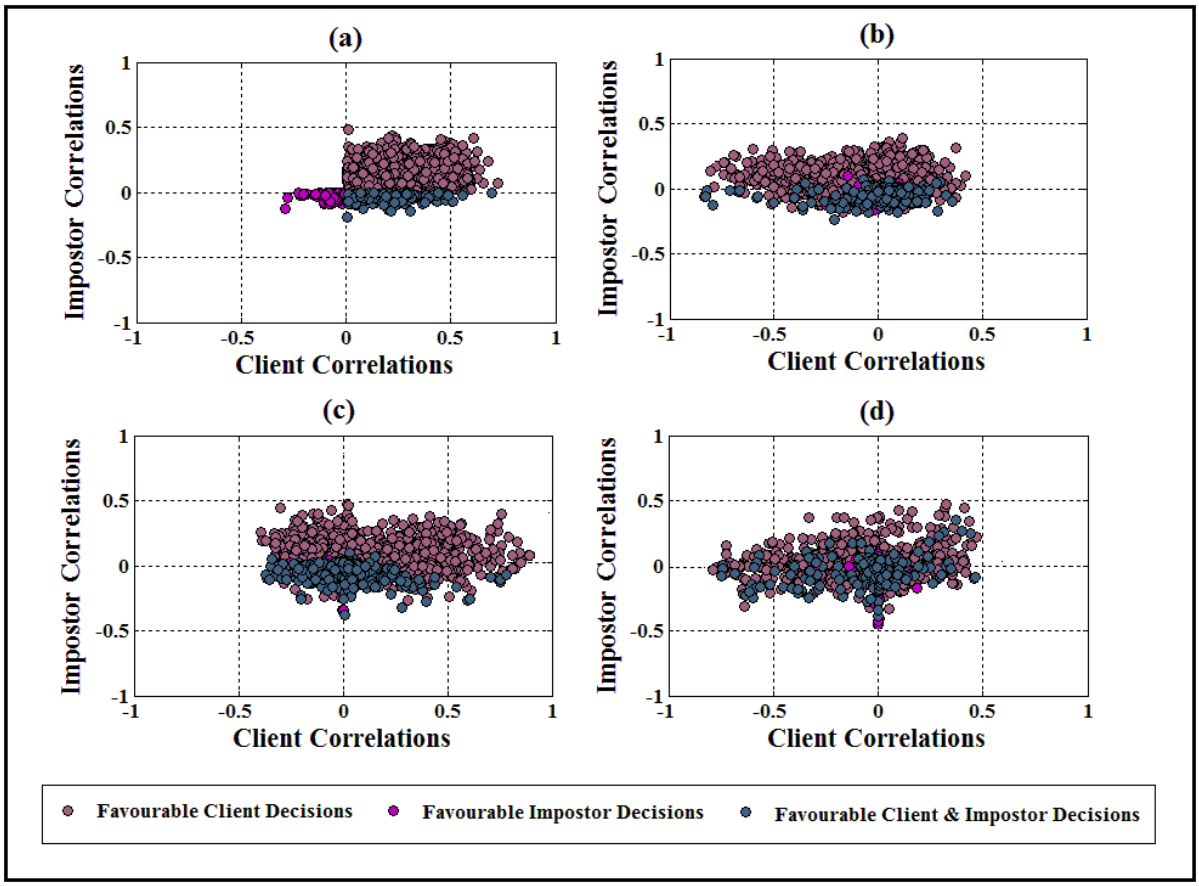

Figure 5.7 Favourable Client and Impostor Correlation Coefficients of (a) 2nd-order (Two-Digit Combinations) (b) 3rd-order (Three-Digit Combinations) (c) 4th-order (Four-Digit Combinations) and (d) 5th-order (Five-Digit Combinations)

combinations) is described in section 3.5.2 for speech data from *SET-1*. In this section, the analysis is presented using the pooled results for all the speakers' test datasets.

Figure 5.7 presents the correlation coefficients that are favourable for client decisions (5.28), impostor decisions (5.26) and client-impostor decisions (5.28 & 5.26) for '*n*' decisions ('*n*' =2, 3, 4 & 5) from speakers of *SET-1*. The client and impostor coefficients are plotted for two-digit combinations (2nd-order - fig. 5.7(a)), three-digit combinations (3rd-order - fig. 5.7(b)), four-digit combinations (4th-order - fig. 5.7(c)) and five-digit combinations (5th-order - fig. 5.7(d)). The positive 2nd-order client correlation coefficients are favourable whereas negative 2nd-order impostor correlation coefficients are favourable. The digits with client correlation coefficients between [0, 1] (positive) and impostor correlation coefficients between [-0.5, 0] (negative) are observed to be favourable for two-digit combinations. The sign of the higher-order coefficients may not be sufficient for determining the favourable

dependence between the decisions. In fig. 5.7, it is demonstrated that the even order coefficients for client decisions are mostly within the range [-0.5, 1] whereas the odd order coefficients are mostly within the range [-1, 0.5]. The even order and odd coefficients for impostor decisions are shown to be in [-0.5, 0.5] for the speakers in *SET-1*. Venkataramani [39] showed that *'AND fusion rule'* is optimal for classifiers with a maximum positive client correlation coefficient of 1 and minimum negative impostor correlation coefficient of -0.5. The range of correlation coefficients for digits with favourable dependence in figure 5.7 also satisfies the conditions, in [39], which represents that *'AND fusion rule'* is optimal for the combination of these digits.

Using (5.28) and (5.26), the combinations with favourable dependence between decisions from multiple instances are determined. The mean error rates (FRR & FAR) for a multi-instance fusion scheme with favourable combinations are presented in fig. 5.8. The total error rates for these digit combinations with favourable dependence are represented in fig. 5.13(a). The error rates are presented for entire set of possible digit combinations and a separate set of digit combinations for the client favourable, impostor favourable and client-impostor favourable combinations that are speaker-specific. The number of digit combinations in each set is different and the mean total error rates are shown to be better when impostor and client-impostor favourable digit combinations are considered. One reason

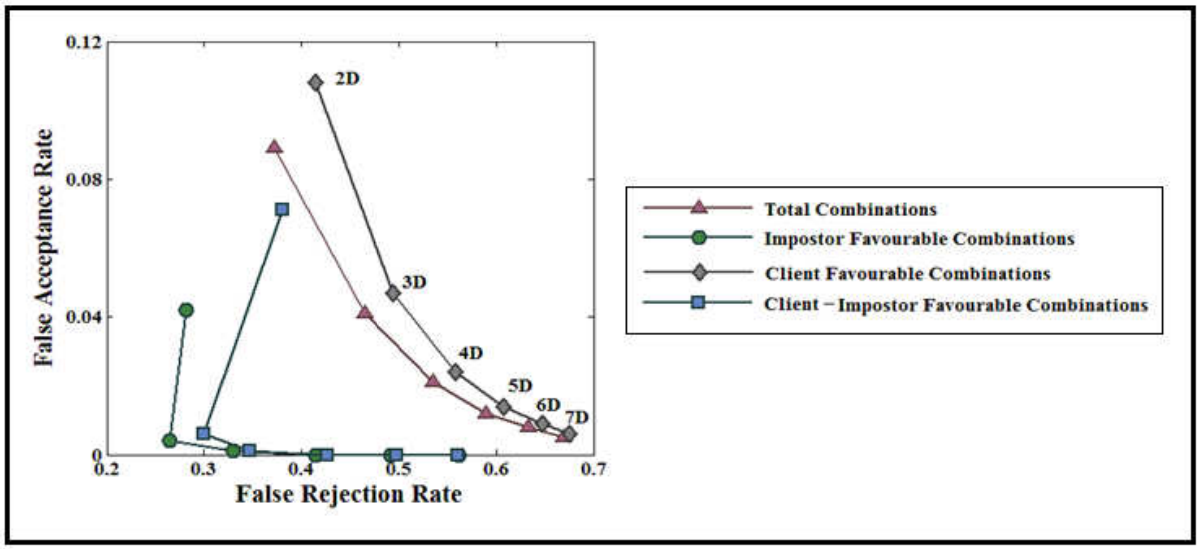

Figure 5.8 Total error rates for multi-instance and multi-sample fusion schemes with client, impostor, client-impostor favourable digit combinations

for this improvement is because of the greater decrease in false acceptance rate with an increase in digits. The figure shows improved performance for the test datasets of *SET-1* but the improvement cannot be expected for all the verification datasets.

In fig. 5.5, each point represents the mean TER for different combinations of *'n'* digits (*n*=1, 2, 3 ... *n*). As the digit combinations at each stage are different, the corresponding theoretical error rates are also different. Table 5.6 presents the ideal error rates calculated under independence assumption and the error rates for the fusion of dependent decisions from multiple instances using *'AND'* fusion. The TER for the fusion of two, four and seven digits are shown as an example. The mean total error rates for ideal case are higher than dependent fusion for entire set with all possible digit combinations. Nevertheless, the difference in these error rates is higher for favourable digit combinations. Although client or impostor favourable combinations have lower total error rates (or TERs), the combinations may not ensure that the individual error rates (FRR and FAR) are with higher *error differences*. Therefore, the client-impostor favourable combinations are preferable for verification of both client and impostors. The errors from this set of combinations always ensure that fusion of *dependent decisions* are always lower than the fusion of *independent decisions* from multiple instances that are combined using *'AND Rule'*.

Table 5.6 Total Error Rates for digit combinations with Favourable Dependence on Client and Impostor Decisions (Ideal - TER for fusion of independent decisions, Exp. - TER for fusion of dependent decisions, *n* - number of digits/instances)

(*n*)	Total Error Rates							
	Client & Impostor		Favourable Client		Favourable Impostor		Favourable Client-Impostor	
	Ideal	Exp.	Ideal	Exp.	Ideal	Exp.	Ideal	Exp.
2	$0.475^{\pm 0.26}$	$0.461^{\pm 0.244}$	$0.547^{\pm 0.238}$	$0.522^{\pm 0.222}$	$0.342^{\pm 0.28}$	$0.322^{\pm 0.26}$	$0.494^{\pm 0.28}$	$0.452^{\pm 0.26}$
4	$0.622^{\pm 0.2}$	$0.556^{\pm 0.212}$	$0.661^{\pm 0.225}$	$0.582^{\pm 0.200}$	$0.352^{\pm 0.20}$	$0.331^{\pm 0.18}$	$0.385^{\pm 0.22}$	$0.348^{\pm 0.19}$
7	$0.771^{\pm 0.23}$	$0.674^{\pm 0.193}$	$0.785^{\pm 0.221}$	$0.681^{\pm 0.192}$	$0.635^{\pm 0.23}$	$0.562^{\pm 0.19}$	$0.646^{\pm 0.24}$	$0.560^{\pm 198}$

5.3.2 Multi-sample Fusion

The analysis on *favourable dependence* enables to determine better relationship between the *correlation coefficients* (for decisions from multiple samples) and the *Error Difference* (the fusion of dependent and independent decisions). The control of 2nd-order coefficients on false reject and false accept error differences is represented in the fig. 5.9 (a) & (b) respectively for three speakers (Spkr-0074, Spkr-0074 & Spkr-0241) from *SET-1*. The negative *Error Difference*, here, represents that ideal error rate is lower than experimental error rate. The 2nd-order client coefficients have strong negative dependence with an error difference in FRR i.e., with the decrease in correlation the error difference in FRR increases. Whereas 2nd-order impostor coefficients have strong positive dependence with an *error difference* in FAR, i.e., the increase in correlation coefficients decreases the error difference of FAR. The results in the fig. 5.7 thus support the conclusion in [38] that negative and positive correlation coefficients are favourable for client and impostor decisions respectively.

From expressions of the correlation coefficient for *'m'* samples (5.7), it is evident that the *m*th order correlation coefficient is dependent on 2^{nd}, 3^{rd} . . . , $(m-1)$th order coefficients (for example (5.5 - 5.7) represents the expressions for 3rd-5th order coefficients in terms of lower order coefficients and base performances). The dependence on lower order combinations results in a weak relationship between error difference and *m*th order

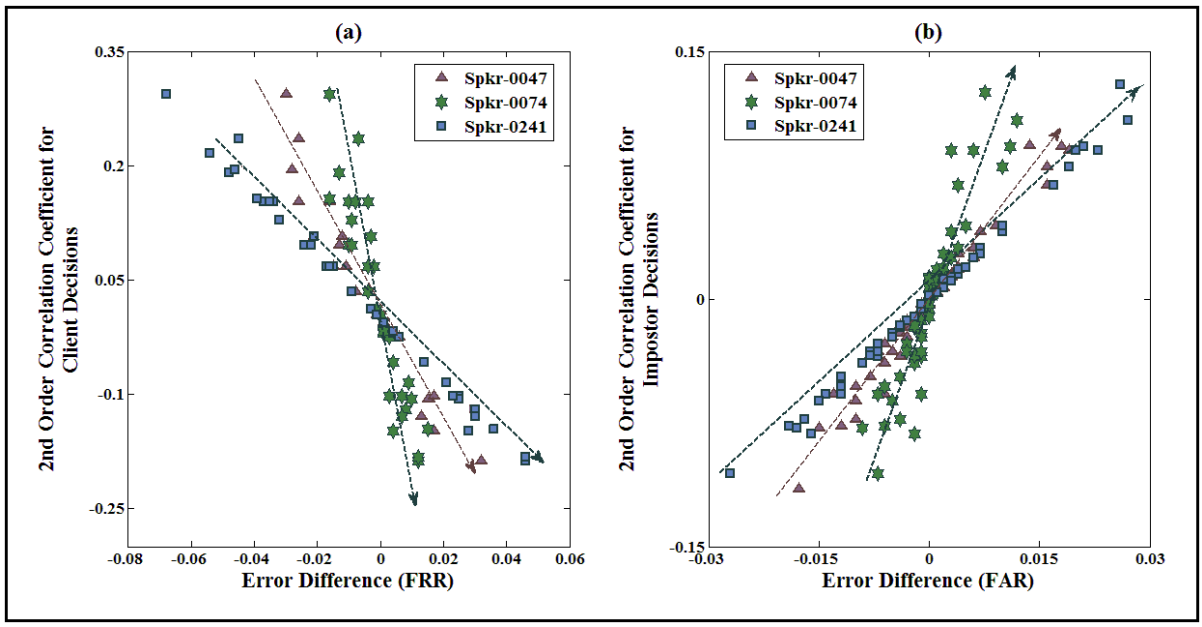

Figure 5.9 Error Differences and 2nd-order Correlation Coefficients for Multi-Sample Fusion of (a) Client Decisions and (b) Impostor Decisions

coefficients. Figure 5.10 (a) and (c) shows this lack of direct relationship between error difference and correlation coefficients (client and impostor) for three-digit combinations of speakers 0074, 0074 and 0241 from *SET-1*. When the dependence of *m*th coefficient on previous order coefficients is relaxed (assumed to be zero), the negative 3rd-order client correlation coefficients and positive 3rd-order impostor correlations are favourable for *'OR Rule'* (fig. 5.10(b) and (d)). The same conclusion is extended for other coefficients (*m*>3) where the favourable dependence is determined under the assumption that the correlation coefficients between 2 to (*m*-1) client and impostor decisions are considered to be zero.

The favourable dependence between '*m*' statistically dependent decisions from multiple samples combined using *'OR Rule'* is analysed using the method similar to multi-instance decision fusion (section 5.3.1). One sufficient condition for favourable dependence

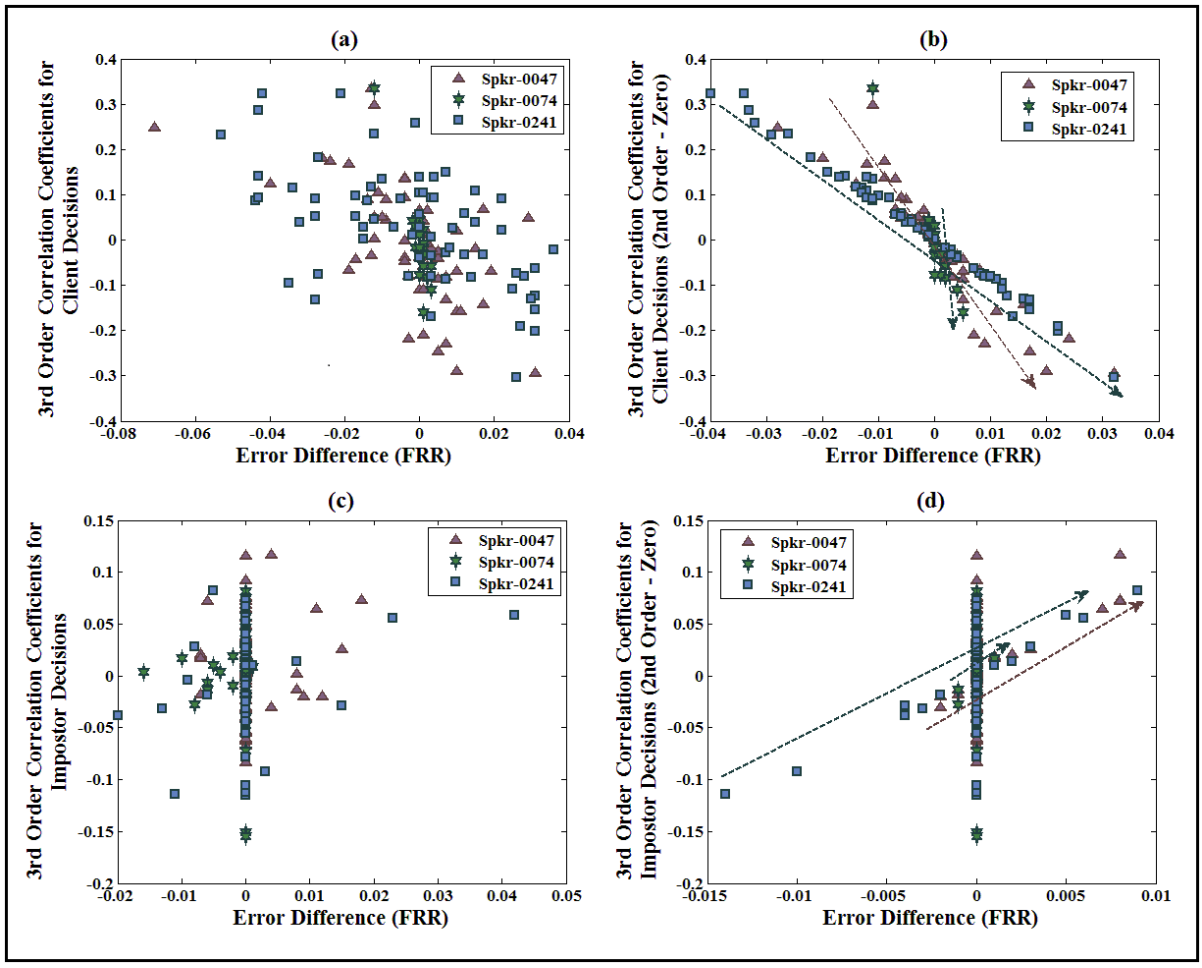

Figure 5.10 Error Differences and 3rd-Order Correlation Coefficients for Multi-Sample Fusion of (a) Three client decisions (b) Three client decisions with 'Zero' 2nd-order coefficients between two client decisions (c) Three impostor decisions and (d) Three impostor

was that the even-order correlation coefficients are negative and odd- order correlation coefficients are positive on client decisions [39]. For impostor decisions, positive even-order correlation coefficients and negative-odd order correlation are favourable [39]. The analysis to determine other conditions of 'favourable/unfavourable' dependence between client and impostor decisions for '*m*' sample fusion using 'OR' fusion rule is similar to *'AND' Rule* (Section 5.3.3 and 5.3.4).

The *'AND fusion rule'* used for combination of multiple instances is the complement of the *'OR fusion rule'* used for combination of multiple samples. Analysis similar to the 'AND' rule can be carried out to find the favourable dependence to the 'OR' fusion rule. Analysis on client decisions for the OR rule is similar to the analysis on impostor decisions for the AND rule. The false rejection rate (FRR) of the multi-sample fusion decreases over the individual classifier FRRs. The generalized equation for favourable dependence between client decisions fused using 'OR' rule is the same as (5.20) in which the error rates and correlation coefficients for impostors are replaced with that of client values, i.e., $\alpha \,\&\, \gamma^0$ are replaced with $\rho \,\&\, \gamma^1$ respectively.

$$\left(\sum_{i<j} \gamma_{ij}^1 \prod_{k=1}^{m} \prod_{k\neq i, k\neq j} \left(\sqrt{\frac{\rho_k}{1-\rho_k}} \right) + \sum_{i<j<k} \gamma_{ijk}^1 \prod_{l=1}^{m} \prod_{l\neq i, i\neq j, i\neq k} \left(\sqrt{\frac{\rho_k}{1-\rho_k}} \right) + .. + \gamma_{12\ldots m}^1 \right) < 0 \qquad (5.29)$$

The 2nd-nth order correlation coefficients of the same sign and negative are favourable whereas coefficients with the same sign and positive are unfavourable. When the coefficients are of different signs, the dependence between the decisions is determined using the base FRR and magnitude of the correlation coefficients.

The false acceptance rate (FAR) of the multi-sample fusion, in general, increases over the individual classifier FRRs. The generalized equation for favourable dependence between impostor decisions combined using 'OR' rule is the same as equation (5.22) in which $\rho \,\&\, \gamma^1$ are replaced with $\alpha \,\&\, \gamma^0$ respectively.

$$\left(\sum_{i<j} \gamma_{ij}^0 \prod_{k=1}^{m} \prod_{k\neq i, k\neq j} \left(\sqrt{\frac{1-\alpha_k}{\alpha_k}} \right) + \sum_{i<j<k} \gamma_{ijk}^0 \prod_{l=1}^{m} \prod_{l\neq i, i\neq j, i\neq k} \left(\sqrt{\frac{1-\alpha_k}{\alpha_k}} \right) + .. + \gamma_{12\ldots m}^0 \right) > 0 \qquad (5.30)$$

The 2nd-nth order correlation coefficients of same sign are favourable when positive and unfavourable when negative. If the coefficients are of different signs, the dependence

between the decisions is determined using the base FAR and magnitude of correlation between impostor decisions, i.e., the sum of (*m*-1)th order correlation coefficients multiplied with FAR factor $\left(\frac{1-\alpha_m}{\alpha_m}\right)$ of the *m*th sample. The ***favourable dependence*** between the decisions from either client or impostors enables to determine the ***best classifier*** that results in lower error rates for fusion of dependent sample decisions rather than fusion of independent decisions using sequential *'OR Rule'*.

5.3.2.1 Error Rates for repeated digit samples with favourable dependence

The analysis on favourable dependence enables to determine the best classifier with *Error Difference* greater than zero (i.e., fusion of dependent decisions from samples results in lower error rates than fusion of independent decisions from repeated samples). The client favourable and impostor favourable combinations for text-dependent speaker verification are evaluated using (5.29) & (5.30) respectively. The protocol used for the evaluation of multi-

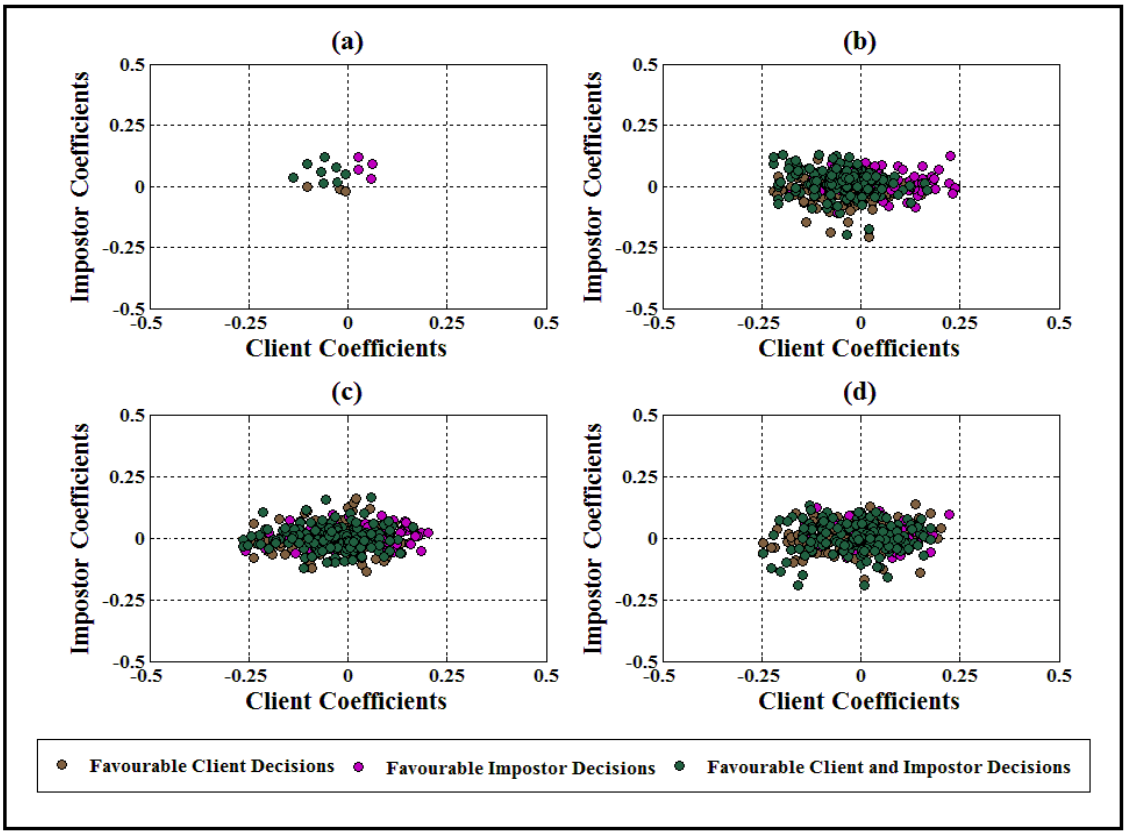

Figure 5.11 Favourable Correlation Coefficients for Client and Impostor Decisions of (a) 2nd-order (b) 3rd-order (c) 4th-order and (d) 5th-order

sample fusion is described in section 3.5.2 for speech data from *SET-1*. In this section, the analysis is presented using the pooled results for all the speakers' test datasets.

Figure 5.11 presents the correlation coefficients that are favourable for client decisions (5.29), impostor decisions (5.30) and client-impostor decisions (5.29 & 5.30) for *'m'* decisions ('*m*' =2, 3, 4 & 5) from speakers of *SET-1*. The client and impostor coefficients are plotted for two-digit samples (2nd-order - fig. 5.11(a)), three repeated sample combinations (3rd-order - fig. 5.11(b)), four sample combinations (4th-order - fig. 5.11(c)) and five sample combinations (5th-order - fig. 5.11(d)). The negative 2nd-order client correlation coefficients are favourable whereas positive 2nd-order impostor correlation coefficients are favourable. The digits with client correlation coefficients between [-0.25, 0] (negative) and impostor correlation coefficients between [0 0.25] (positive) are observed to be favourable for fusion of two samples. Venkataramani [39] showed that *'OR fusion rule'* is optimal for classifiers with a positive correlation coefficient of impostor scores with the maximum correlation coefficient of 1 and negative correlation coefficient of client scores with a minimum correlation coefficient of -0.5. The range of correlation coefficients for digits with favourable dependence in figure 5.5 also satisfies the condition, in [39], which represents that *'OR fusion rule'* is optimal for the combination of repeated samples of a digit.

Using (5.29) and (5.30), the digits with favourable dependence between decisions from multiple samples are determined. Figure 5.12 presents the error rates (FRR & FAR) for

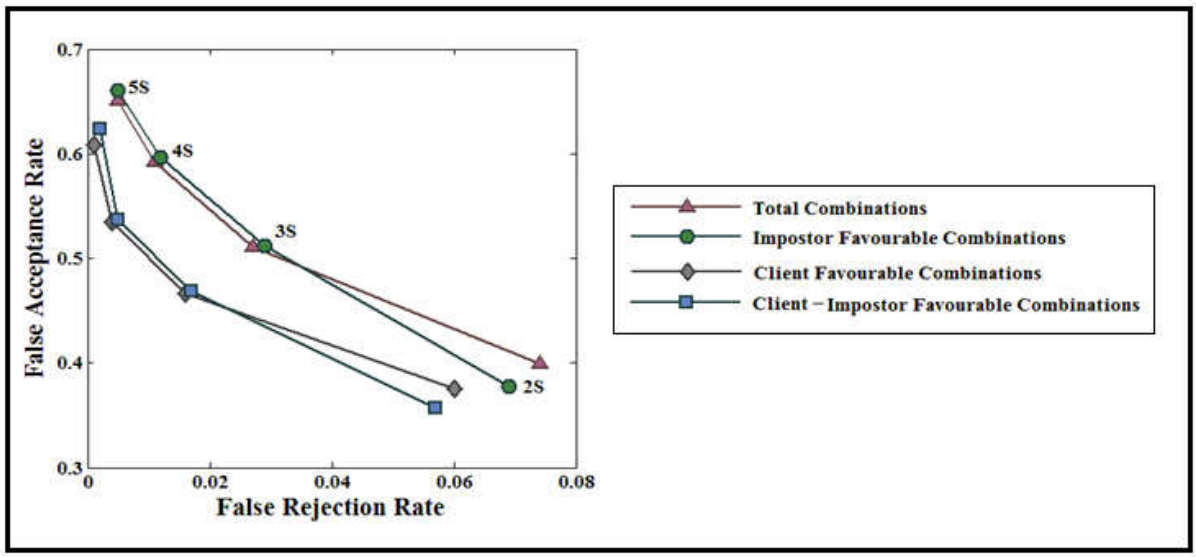

Figure 5.12 Total error rates for multi-instance and multi-sample fusion schemes with client, impostor, client-impostor favourable digit combinations

multi-sample fusion with favourable digits. The error rates are presented for all the digits in dataset and a separate set of digits with favourable dependence for client, impostor and client-impostor samples that are speaker-specific. The number of favourable digits in each set is different and the mean total error rates are shown to be better when impostor and client-impostor favourable digit combinations are considered. One reason for this improvement is the greater decrease in the false rejection rate with an increase in samples used for verification. The figure shows improved performance for the test datasets of *SET-1* but the improvement cannot be expected for all the verification datasets.

Table 5.7 presents the ideal error rates calculated under independence assumption and the error rates for the fusion of dependent decisions from multiple sample using *'OR fusion'*. The total error rates for the ideal case are higher than dependent fusion but the difference in error rates are observed to increase for the use of favourable digits for testing. Although, the use of client or impostor favourable digits for dependent decisions result in lower overall

Table 5.7 Total Error Rates for decisions from Digits with Favourable Dependence for Client and Impostor Correlation Coefficients (Ideal TER-Ideal Total Error Rate, Exp. TER-Experimental Total Error Rate ; 2S -Two Samples, 3D-Three Samples, 4D-Four Samples and 5D-Five Samples)

	Client & Impostor Coefficients $(\gamma^0 \& \gamma^1)$		Favourable Client-Impostor Coefficients		Favourable Client Coefficients		Favourable Impostor Coefficients	
	Ideal TER	Exp. TER	Ideal TER	Exp. TER	Ideal TER	Exp. TER	Ideal TER	Exp. TER
2S	$0.476^{\pm 0.3}$	$0.475^{\pm 0.3}$	$0.514^{\pm 0.3}$	$0.496^{\pm 0.3}$	$0.436^{\pm 0.3}$	$0.427^{\pm 0.3}$	$0.514^{\pm 0.3}$	$0.496^{\pm 0.3}$
3S	$0.540^{\pm 0.3}$	$0.538^{\pm 0.3}$	$0.525^{\pm 0.3}$	$0.506^{\pm 0.3}$	$0.498^{\pm 0.3}$	$0.488^{\pm 0.3}$	$0.563^{\pm 0.3}$	$0.552^{\pm 0.3}$
4S	$0.603^{\pm 0.3}$	$0.603^{\pm 0.3}$	$0.554^{\pm 0.3}$	$0.542^{\pm 0.3}$	$0.541^{\pm 0.3}$	$0.537^{\pm 0.3}$	$0.623^{\pm 0.3}$	$0.614^{\pm 0.3}$
5S	$0.658^{\pm 0.3}$	$0.657^{\pm 0.3}$	$0.658^{\pm 0.3}$	$0.647^{\pm 0.3}$	$0.642^{\pm 0.3}$	$0.640^{\pm 0.3}$	$0.679^{\pm 0.3}$	$0.670^{\pm 0.3}$

error rate (or TER), these combinations may not ensure that the individual error rates (FRR and FAR) are with higher *error differences*. Therefore, the client-impostor favourable combinations are preferable for verification of both client and impostors as the fusion of dependent decisions for these combinations are better than the fusion of independent decisions.

5.3.3 Multi-instance and Multi-sample Fusion ('*n*' instances and '*m*' samples)

The estimation of error rates for multi-instance and multi-sample combinations using correlation modelling are explained in the sections 5.3 and 5.4 respectively. The decision correlation for the combination of multiple instances with single sample and multiple samples are different. This analysis is explained by determining the favourable dependence between impostor decisions for the proposed system.

The *'OR rule'* is used to determine if the speaker is accepted at a decision stage (an instance level). Considering $\alpha_{S1}^{C1} \& \alpha_{S2}^{C1}$ being the FAR for the samples *S1* and *S2* respectively for a classifier/instance *C1*, the error rates for the two-sample fusion of an instance can increase the false accepts. The nature of the repeated sample here could be a random where $\alpha_{S1}^{C1} = \alpha_{S2}^{C1}$ or adaptive where $\alpha_{S1}^{C1} \le \alpha_{S2}^{C1}$. The false accepts for the fusion of these multiple samples '*S1* & *S2*' can be expressed using BL expansion. The expressions for the fusion of two samples for instances *C1* & *C2* are given as:

$$\alpha_{S1,S2}^{C1} = 1 - \left(\left(1-\alpha_{S1}^{C1}\right)\left(1-\alpha_{S2}^{C1}\right)\right)\left(1+\gamma_1^1\sqrt{\frac{\alpha_{S1}^{C1}\,\alpha_{S2}^{C1}}{\left(1-\alpha_{S1}^{C1}\right)\left(1-\alpha_{S2}^{C1}\right)}}\right)$$
$$\alpha_{S1,S2}^{C2} = 1 - \left(\left(1-\alpha_{S1}^{C2}\right)\left(1-\alpha_{S2}^{C2}\right)\right)\left(1+\gamma_2^1\sqrt{\frac{\alpha_{S1}^{C2}\,\alpha_{S2}^{C2}}{\left(1-\alpha_{S1}^{C2}\right)\left(1-\alpha_{S2}^{C2}\right)}}\right) \quad (5.31)$$

The claim is declared genuine, at the end of *'n'* decision stages (or instances), if a speaker is accepted at all the instances. Hence, '*AND Rule'* is used to determine the acceptance between the instances. Considering $\alpha_{S1,S2}^{C1} \& \alpha_{S1,S2}^{C1}$ being the FAR for the instances *C1* and *C2* respectively with two samples for each instance, the error rates for the combination of two

instances can decrease the number of false accepts $(\alpha_{S1,S2}^{C1,C2})$. The expression for the FAR of multi-instance fusion (*n*=2) is given as:

$$\alpha_{S1,S2}^{C1,C2} = \alpha_{S1,S2}^{C1}\ \alpha_{S1,S2}^{C2}\left(1+\gamma_{12}^{1}\sqrt{\frac{\left(1-\alpha_{S1,S2}^{C1}\right)\left(1-\alpha_{S1,S2}^{C2}\right)}{\alpha_{S1,S2}^{C1}\ \alpha_{S1,S2}^{C2}}}\right) \tag{5.32}$$

For the condition of favourable dependence between the decisions, the difference between ideal and experimental error rates is to be greater than zero. The inequality for favourable dependence is given as

$$\gamma_{12}^{1}\sqrt{\alpha_{S1,S2}^{C1}\ \alpha_{S1,S2}^{C2}\left(1-\alpha_{S1,S2}^{C1}\right)\left(1-\alpha_{S1,S2}^{C2}\right)} < 0 \tag{5.33}$$

As FARs have the values between zero and one (inclusive), the condition satisfying the above condition needs the correlation coefficient for impostor decisions to be negative.

The analysis of favourable dependence for sequential decision fusion of *n*=3 is similar to that of multi-instance fusion. The dependence equation for instances with different error rates is expressed as

$$\left(\gamma_{12}^{1}\sqrt{\left(\frac{\alpha_{S1,S2}^{C3}}{1-\alpha_{S1,S2}^{C3}}\right)}+\gamma_{13}^{1}\sqrt{\left(\frac{\alpha_{S1,S2}^{C2}}{1-\alpha_{S1,S2}^{C2}}\right)}+\gamma_{23}^{1}\sqrt{\left(\frac{\alpha_{S1,S2}^{C1}}{1-\alpha_{S1,S2}^{C1}}\right)}+\gamma_{123}^{1}\right) < 0 \tag{5.34}$$

When the 2nd-order and 3rd-order correlation coefficients are of the same sign and positive, the correlation factor in (5.34) is positive and therefore the condition for favourable dependence is not satisfied. If the 2nd-order and 3rd-order correlation coefficients are of same sign and negative, the (5.34) is justified and therefore is a satisfying condition for favourable dependence. Thus, the negative 2nd and 3rd-order coefficients are favourable whereas positive 2nd and 3rd-order coefficients are unfavourable. When the 2nd and 3rd-order coefficients are of different signs, the condition for determining favourable dependence depends on the sign of correlation between the instances, coefficients between the samples for each instance and base FAR of instances.

The analysis for favourable dependence of decisions from *n* instances is similar to that of three instances. The generalised condition for determining the favourable dependence with individual FAR $\alpha^{Ci}_{S1,S2,..Sm}(i=1,2,3...n)$ for '*n*' instances with '*m*' samples is given as

$$\left(\sum_{i<j}\gamma^{1}_{ij}\prod_{k=1}^{n}\prod_{k\neq i,k\neq j}\left(\sqrt{\frac{\alpha^{Ck}_{S1,S2,..Sm}}{1-\alpha^{Ck}_{S1,S2,..Sm}}}\right)+\sum_{i<j<k}\gamma^{1}_{ijk}\prod_{l=1}^{n}\prod_{l\neq i,l\neq j,l\neq k}\left(\sqrt{\frac{\alpha^{Cl}_{S1,S2,..Sm}}{1-\alpha^{Cl}_{S1,S2,..Sm}}}\right)+..+\gamma^{1}_{123...n}\right)<0 \qquad (5.35)$$

Where $\alpha^{Cn}_{S1,S2,S3.....,Sm}$ is expressed using (5.16). Thus the determination of dependence for the fusion of '*m*' samples and '*n*' instances depend on the base performances, correlation coefficients between the repeated samples and correlation between the instances being combined. Due to complex relationship between the terms and non-linearity of multiple correlation coefficients, the solution is intractable and difficult to solve.

The solution for favourable dependence between client decisions is analysed in steps similar to that of impostor decisions. If the speaker is rejected for all the repeated samples of an instance, the claim by the client is rejected. The false rejects for the fusion of '*m*' multiple samples can thus be determined using the *'AND' logic*. If the speaker is rejected at any instance, the client claim is rejected. The false rejects for the fusion of '*n*' multiple instances can thus be determined using the 'OR' logic. The generalised equation for determining the favourable dependence with individual FRR $\rho^{Ci}_{S1,S2,..Sm}(i=1,2,3...n)$ for '*n*' instances with '*m*' samples is given as

$$\left(\sum_{i<j}\gamma^{0}_{ij}\prod_{k=1}^{n}\prod_{k\neq i,k\neq j}\left(\sqrt{\frac{1-\rho^{Ck}_{S1,S2,..Sm}}{\rho^{Ck}_{S1,S2,..Sm}}}\right)+\sum_{i<j<k}\gamma^{0}_{ijk}\prod_{l=1}^{n}\prod_{l\neq i,l\neq j,l\neq k}\left(\sqrt{\frac{1-\rho^{Cl}_{S1,S2,..Sm}}{\rho^{Cl}_{S1,S2,..Sm}}}\right)+..+\gamma^{0}_{123...n}\right)>0 \quad (5.36)$$

When the 2nd-nth order impostor correlation coefficients are of same sign and negative, the dependence is favourable (5.35) whereas positive 2nd-nth correlation coefficients are favourable on client decisions (5.36). For correlation coefficients with different signs, favourable dependence can be determined between the samples for each instance and base error rates of the instances. The above analysis of favourable dependence for the proposed fusion enables to find instance combinations with experimental/predicted error rates smaller than ideal error rates.

5.3.3.1 Error Rates for favourable digit combinations

The correlation between decisions from multiple instances and multiple samples enables to define an accurate relationship between the experimental and ideal error rates. Table 5.8 presents the correlation coefficients that are favourable for client decisions, impostor decisions, and client and impostor decisions for speakers from *SET-1*. The favourable dependence for client and impostor decisions are calculated using (5.35) and (5.36) for multi-instance and multi-sample fusion schemes ('*n*' =2, 3 ... 7). From table 5.8, the negative 2nd-order coefficients are shown to be favourable for impostor decisions whereas positive 2nd-order coefficients are shown favourable for client decisions. There exists no direct relationship between error difference and higher-order coefficients (>2). However, it is shown (explained using 5.35 & 5.36) that correlations of same sign and positive are mostly favourable for client decisions whereas impostor correlations of same sign and negative are favourable. However, this dependence on previous order coefficients of multiple instances and multiple samples can be relaxed in the case where lower order coefficients are

Table 5.8 Correlation Coefficients for digit combinations with multiple samples for favourable client and impostor decisions (2D-1S: Two Digits - One Sample, 2D-2S: Two Digits - Two Samples ...)

	Mean Client and Impostor Coefficients		Favourable Client and Impostor Coefficients		Favourable Client Coefficients		Favourable Impostor Coefficients	
	Client	Impostor	Client	Impostor	Client	Impostor	Client	Impostor
2D-1S	$0.169^{\pm 0.20}$	$0.111^{\pm 0.10}$	$0.221^{\pm 0.16}$	$-0.037^{\pm 0.0}$	$0.257^{\pm 0.16}$	$0.125^{\pm 0.10}$	$0.080^{\pm 0.18}$	$-0.033^{\pm 0.}$
2D-2S	$0.051^{\pm 0.15}$	$0.052^{\pm 0.07}$	$0.169^{\pm 0.13}$	$-0.032^{\pm 0.0}$	$0.164^{\pm 0.14}$	$0.049^{\pm 0.07}$	$0.054^{\pm 0.15}$	$-0.034^{\pm 0}$
3D-2S	$0.005^{\pm 0.15}$	$0.008^{\pm 0.06}$	$0.023^{\pm 0.15}$	$-0.035^{\pm 0.03}$	$0.012^{\pm 0.18}$	$0.006^{\pm 0.06}$	$0.013^{\pm 0.12}$	$-0.035^{\pm 0}$
4D-3S	$-0.004^{\pm 0.}$	$0.003^{\pm 0.06}$	$-0.044^{\pm 0.19}$	$-0.018^{\pm 0.04}$	$-0.005^{\pm 0}$	$-0.002^{\pm 0}$	$-0.036^{\pm 0}$	$-0.017^{\pm 0}$
5D-4S	$-0.002^{\pm 0}$	$-0.002^{\pm 0.05}$	$-0.002^{\pm 0.01}$	$-0.009^{\pm 0.05}$	$-0.003^{\pm 0.}$	$-0.001^{\pm 0.}$	$-0.001^{\pm 0.}$	$-0.009^{\pm 0}$

considered zero. Further, the correlation values decrease with the increase in either the number of samples or instances used for fusion. The correlation values are often undefined for the case where the error rates are zero because of the insufficient number of tests performed.

Figure 5.13 presents the mean total error rates (TER) for digit combinations that are favourable for client, impostor and client-impostor decisions. The correlation combinations

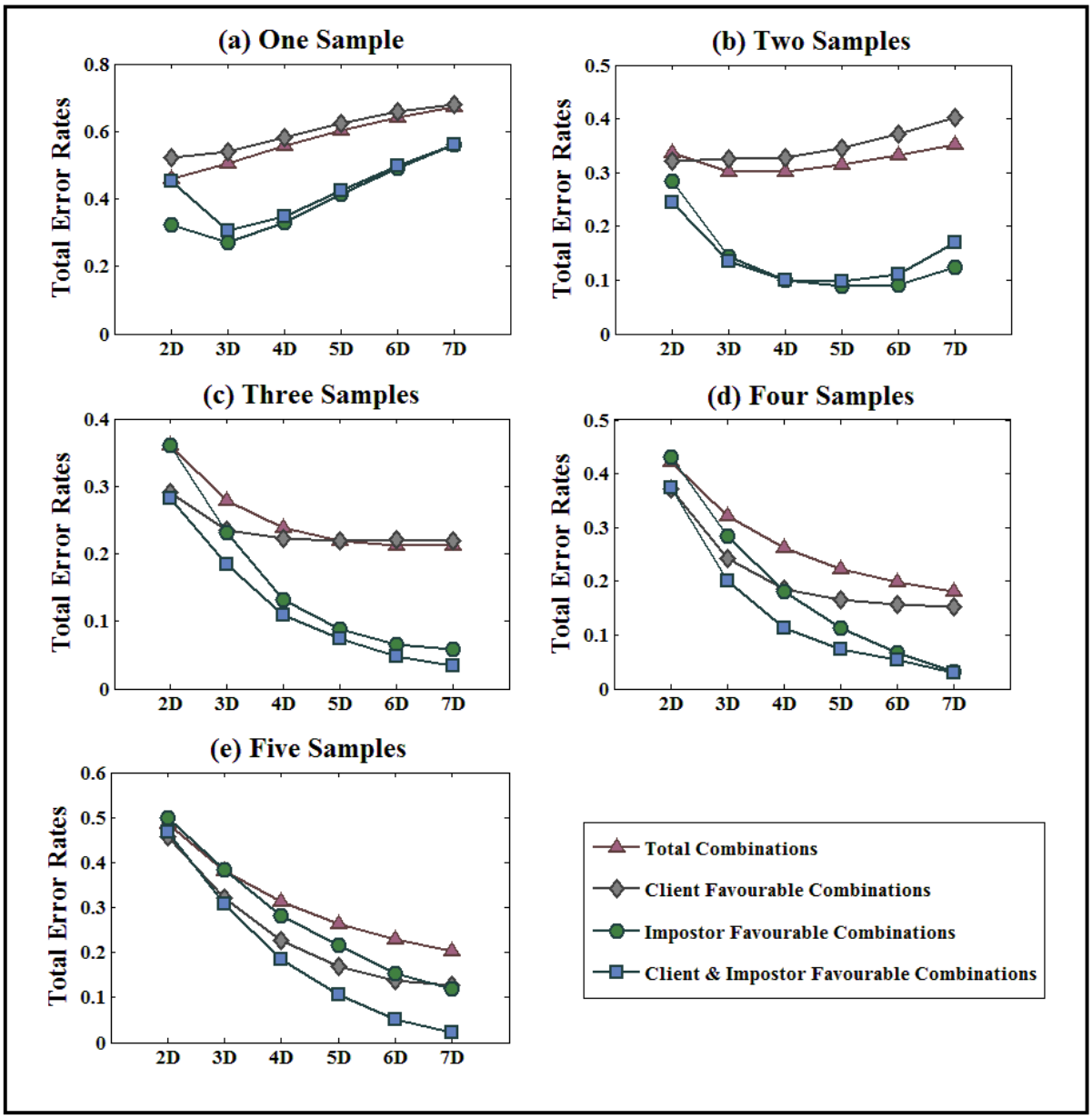

Figure 5.13 Total Error Rates for multi-instance and multi-sample fusion schemes with favourable dependence for Client, Impostor, Client & Impostor decisions for two digits (2D), three digits (3D), four digits (4D), five digits (5D), six digits (6D) and seven digit

represented in fig. 5.13 are shown in table 5.8. The TER decreases with an increase in samples initially and then progressively increases when the increase in FAR is higher than decrease in FRR for the use of multiple samples (e.g., fusion of five samples results in TER higher than fusion of four samples). The same could be explained for the case where the TER for digit combinations with favourable impostor decisions. When the favourable digit combinations for client and client-impostor decisions are considered, the TER decreases progressively for each additional sample. The TER of 20.2% for digit combinations with parameters (7-instances, 5-samples) are reduced to 12.6% and 11.9% when digit combinations only with favourable client and impostor decisions respectively are selected. The verification based on digit combinations that are favourable for both client and impostor

Table 5.9 Verification error rates for multi-instance and multi-sample fusion schemes with favourable dependence for client, impostor, client & impostor decisions ('n' - number of instances and 'm' - number of samples)

Errors (in %)	*(n, m)*	Total Combinations		Client-Impostor Favourable		Client Favourable		Impostor Favourable	
		Ideal	Exp	Ideal	Exp	Ideal	Exp	Ideal	Exp
FRR	(1, 1)	23.7	23.7	23.7	23.7	23.7	23.7	23.7	23.7
	(2, 1)	40.5	37.2	41.7	38.1	46.2	41.4	29.6	28.1
	(3, 2)	19.8	18.7	9.6	8.9	21.8	19.9	10.3	9.9
	(4, 3)	10.1	9.7	5.1	4.8	9.5	8.9	6.0	5.8
	(5, 4)	5.4	5.3	2.1	2.0	3.9	3.7	3.1	3.0
	(6, 5)	6.4	2.7	2.0	0.6	5.1	1.4	2.6	2.0
FAR	(1, 1)	23.8	23.8	23.8	23.8	23.8	23.8	23.8	23.8
	(2, 1)	7.0	8.9	7.7	7.1	8.5	10.8	4.6	4.2
	(3, 2)	9.8	11.4	4.7	4.5	11.2	12.7	4.8	4.6
	(4, 3)	12.9	14.1	6.4	6.2	12.2	13.3	7.6	7.3
	(5, 4)	16.1	17.1	5.7	5.4	12.0	13.0	8.6	8.3
	(6, 5)	19.4	20.1	4.6	4.4	11.6	12.4	13.6	13.3

decisions can reduce the TER to 2.1%. Further, TER is reduced in most cases with the increase in instances used for fusion. The fusion of 5-instances and 5-samples results in TER of 10.7% whereas the combination of six instances with five samples for each instance reduces the TER to 5% (FRR - 0.6% and FAR - 4.4%).

Table 5.9 presents the verification error rates for fusion of independent and dependent decisions (with favourable dependence for client, impostor and client & impostor decisions) for multiple instances and samples. The false rejection and false acceptance rates for fusion of client-favourable decisions are always lower than that of independent decisions. However, the false rejection rates and false acceptance rates for fusion of impostor favourable decisions and client-favourable decisions may not always be lower than that of independent decisions. From the digit decisions from *SET-1*, it is shown that the experimental FRRs are lower than ideal/independent FRRs for impostor favourable digit combinations FRRs whereas for client favourable digit combinations the ideal FARs are lower than experimental FARs.

The favourable digit combinations are different between speakers whereas similar for the same speaker across different datasets thereby ensuring an accurate final decision on the identity claim. For example, the digits in sequence 2-9-1-3-7, i.e., the combinations 29, 291, 2913, 29137 are favourable for Spkr-0047 whereas the sequence 2-3-1-4-7 is favourable for Spkr-0241 across datasets. Nevertheless, the sequence 2-5-3-4-7, which is unfavourable for Spkr-0047, is observed to be favourable for Spkr-0241. An observation here is that the number of combinations with favourable dependence for client-impostor decisions increases with samples used for fusion.

The false acceptance rate for multi-instance fusion is lowered with addition of instances whereas the false rejection rate is reduced by the increase in repeated samples used for fusion. However, with each increase in an instance or sample, the number of correlation coefficients required for accurate prediction of error rates increases exponentially thereby increasing computational cost. It is therefore significant to determine the limit on order of correlation coefficients required for prediction of false acceptance and false rejection rates.

5.4 Limit on the order of correlation for error prediction

The performance gain achieved using multi-instance fusion are initially monotonic but soon reaches saturation, using more instances of the same biometric trait cannot help improve the performance further [213]. With the increase in instances, the number of

combinations increases, i.e., for the fusion of seven digits, there are 21 two-digit combinations, 35 three-digit combinations, 35 four-digit combinations, 21 five-digit combinations and 7 six-digit combinations and therefore the same number of correlation coefficients required for estimating the error rates for a seven-digit combination. In the most general case, $2\left(2^n - n - 1\right)$ correlation coefficients are required in order to obtain the optimal likelihood ratio [209]. As it is clearly impossible to compute the correlation coefficients for an exponential number of combinations, the simplified formulation is proposed where only the most important second and third-order correlations can be used [214]. Losee [215] suggested the truncation of higher-order coefficients (>3) because the incorporation of more dependence information results in a relatively little increase in performance. However, an injudicious truncation of the series may produce unreliable results [216]. For example, the 2nd-order coefficients become negative when the decisions for pairs of instances are close to zero, but the individual error rates are positive; this may lead to the computation of negative error rates from the expansion when third and higher-order dependencies are neglected. In [209] equations are derived to determine approximately the range over which the third-order correlation coefficient can be neglected. This range is estimated based on the ratio between the 3rd and 2nd-order correlation. In most cases, however, the higher order correlation coefficients are neglected when they are found to be very small and consequently has a low effect on the estimation of final error values [196].

The higher-order correlation values for speaker verification decisions shown in table 5.8 are very low and decreases with each progressive addition of a digit. Table 5.10 shows the comparison of FRR and FAR for seven-digit combination with and without neglecting correlations of higher order for client and impostors. The difference in error rates is small when 5th-7th order coefficients are neglected. The error rates are lowered when the correlation among multiple samples is considered. It is evident in the table that only 2nd and 3rd-order coefficients are of importance and thus result in significant error difference for the case of with and without neglecting 3rd-7th order correlations when multiple samples (two and three samples) are used for seven-digit combination. The use of only 2nd and 3rd-order coefficients enables the simplification of BLE expansion in estimation of errors with less complexity. The order of correlation that can be neglected depends on the difference between errors estimated with and without correlation coefficients neglected. The permissible difference between the errors is application dependent and number of digits (instances) used for fusion.

Table 5.10 Error Rates for digit combination with non-zero and zero higher order coefficients

	Samples	2nd-7th order correlations	Zero 7th order correlations	Zero 6th - 7th order correlations	Zero 5th - 7th order correlations	Zero 4th - 7th order correlations
FRR	1	$0.669^{\pm 0.19}$	$0.668^{\pm 0.19}$	$0.671^{\pm 0.19}$	$0.663^{\pm 0.19}$	$0.708^{\pm 0.20}$
	2	$0.478^{\pm 0.14}$	$0.478^{\pm 0.14}$	$0.478^{\pm 0.14}$	$0.479^{\pm 0.14}$	$0.479^{\pm 0.14}$
	3	$0.382^{\pm 0.05}$	$0.382^{\pm 0.05}$	$0.382^{\pm 0.05}$	$0.382^{\pm 0.05}$	$0.381^{\pm 0.05}$
FAR	1	$0.005^{\pm 0.005}$	$0.005^{\pm 0.005}$	$0.005^{\pm 0.005}$	$0.004^{\pm 0.001}$	$0.002^{\pm 0.001}$
	2	$0.022^{\pm 0.02}$	$0.022^{\pm 0.02}$	$0.021^{\pm 0.02}$	$0.021^{\pm 0.02}$	$0.020^{\pm 0.02}$
	3	$0.058^{\pm 0.05}$	$0.058^{\pm 0.05}$	$0.058^{\pm 0.05}$	$0.058^{\pm 0.05}$	$0.057^{\pm 0.05}$

The evaluation results for the effects of favourable dependence and the limits on an order of correlation coefficients are performed on test datasets where the data is known. However, in real scenario, the data is unknown and thus estimating the error rates requires base performances and an estimate on correlation between client and impostor decisions for a speaker. In chapter 4, the error rates for unknown data are assumed equal to base performances of known data when the decisions are independent. The estimation of error rates for correlated decisions is based on BLE where the base performances and correlation between decisions of instances/samples used for verification. The next section deals with the method to estimate the error rates (FRR & FAR) when the speaker data is unknown.

5.5 Estimation of error rates using 'Evaluation and Selection' method

OVERVIEW: In this section, the Selection and Evaluation method is used for the determination of verification error rates for the proposed fusion scheme. The favourable combinations specific for an individual are pre-determined on the development set and later used in speaker testing for test dataset. This method of evaluation is shown to better predict the predicted fusion parameters, i.e., the number of digits and samples or variance of correlation, used for verification always produce a reliable final decision and with errors

lower than that of errors estimated with ideal condition of statistical independence between decisions.

The expressions derived above for error estimation or favourable dependence can be used to tune the parameters, such as the number of instances, the number of samples and the favourable set of digit sequences, limit on the order of the correlation coefficient, required to determine the performance of the fusion method on test data set (unknown data). For tune/development datasets, the correlations between decisions are known and so the experimental values obtained are equal to the estimated values obtained using (5.35) & (5.36). However, in real-world applications (unknown dataset), the correlation values are unknown. In order to estimate the error rates for the test set, the correlation coefficient for a speaker across different tune datasets can be used. The calculation of correlation coefficients, however, depends on the base performances. The base performances are varied based on mismatch between the tune/development and test datasets. The error rates are also varied for different datasets for the same speaker. It is, therefore, significant to determine the influence of variations in performance and thus correlation on the fusion performance of unknown data.

The protocol for performance evaluation is similar to that of multi-instance fusion (section 3.5.2) uses speech data for Spkr-0047 from *SET-1*. As the correlation values are dependent both on the speaker and the digit combination, the results are presented here for only one speaker. The error rates for the test dataset are estimated using the correlation coefficients for a speaker across different tune datasets that consider all the (prior) conditions under which a speaker may be tested.

Figure 5.14 (a) & 5.14 (b) shows the mean correlation coefficients (2nd-order) for tune and test datasets for two-digit combinations of client and impostor decisions respectively. It is noted that there is an overlap between the correlation sets for tune and test datasets. The higher-order correlation coefficients (greater than two) for test dataset are also estimated in the similar manner. The test-dataset error rates can be calculated using the error rates for base classifier and the variance of correlation coefficients for the tune dataset for a specific speaker. The variance in correlation allows the estimation of maximum and minimum error rates (i.e., error bounds) for fixed '*M*' and '*N*' values using derived equations with individual error rates for each instance from tune dataset. The error rates estimated based on the variances of correlation coefficients can result in maximum and the minimum error bounds. The error bounds obtained using the correlation coefficients for two-digit

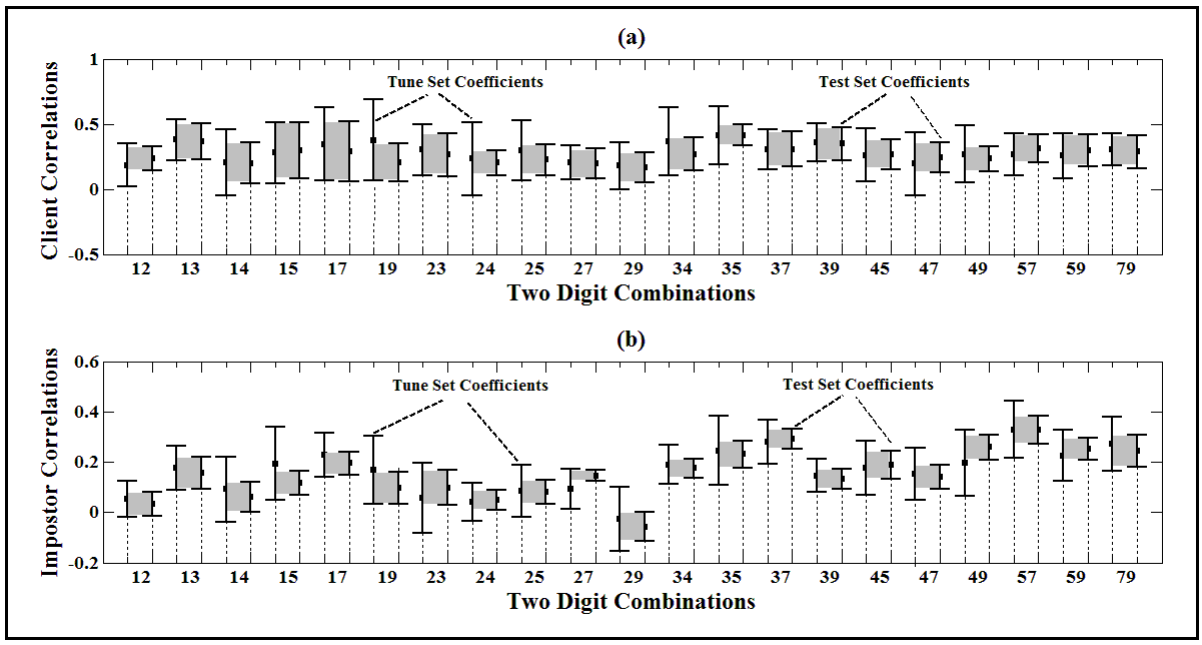

Figure 5.14 Mean 2nd-order Correlation Coefficients for Tune and Test Datasets for Speaker-0047 in *SET-1* (a) client correlations and (b) impostor correlations

combinations are shown in fig. 5.15 (a) & 5.15 (b). It is evident that most of the experimental error rates for the test dataset fall within the bounds of error rates estimated using the maximum and minimum correlation coefficients for each two-digit combination from tune dataset. The digit combinations greater than two (n>2) also has the experimental error rates between the estimated error bounds.

The error rates for base classifiers and the variance in correlation coefficients from the tune dataset (known data) are used to estimate the fusion performance of test dataset (unknown data). As the base classifiers for both the tune and test datasets are similar (no mismatch between tune and test set conditions), the parameters (n, m) used to control the trade-off between FRR and FAR on tune dataset are applied to test dataset also. The fusion error rates for these two datasets can be expected to have slight difference because of the variations in the decision correlation. This difference in error rates can be used as another measure to determine the identity claim. If the difference in the error rates for digit combinations is high then there is higher possibility that the unknown data being tested might not belong to the claimed speaker.

The theoretical equations can also be used to estimate the parameters required to obtain desired performance for a verification system. In real-world applications, the verification system may set initial acceptable values for FRR and FAR. The derived

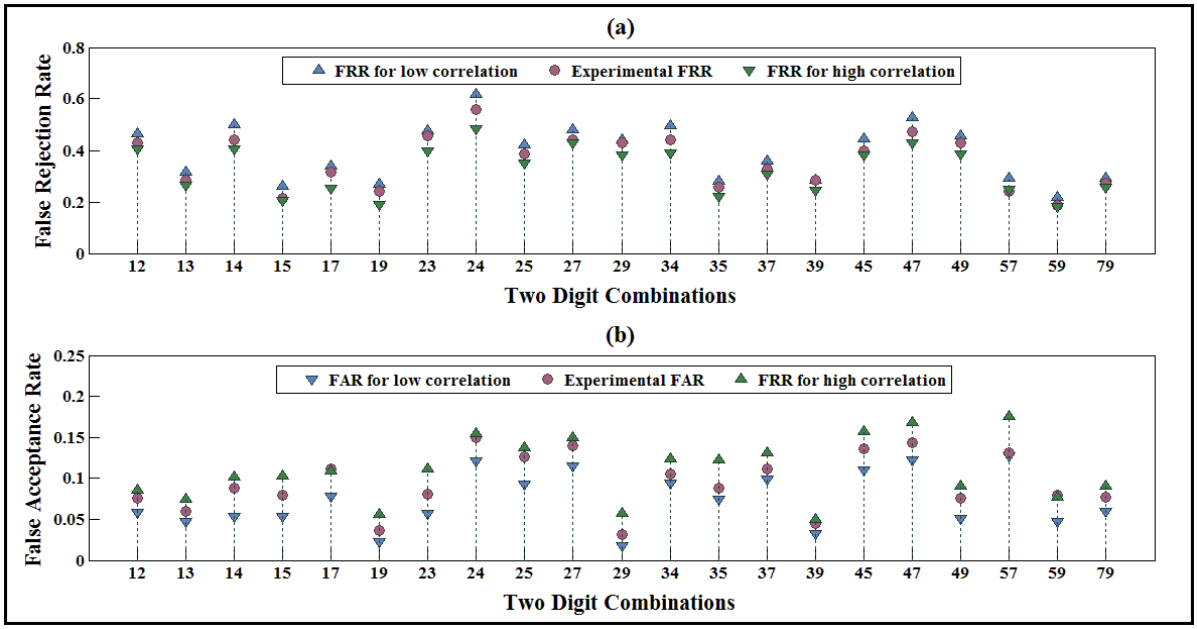

Figure 5.15 Experimental and predicted error rates for test dataset of Speaker-0047 from *SET-1* (a) False Rejection Rate and (b) False Acceptance Rate

equations are used to estimate the number of instances and samples required to obtain the experimental error rates that are equal or lower than the desired FRR and FAR for a speaker. The other parameters required for this estimation are the base error rates and the variance in correlation coefficients of the known data. These parameters required to obtain the desired performance are user dependent.

5.6 Chapter Summary and Conclusion

The architecture for sequential *'AND'* fusion of instances and sequential *'OR'* fusion of samples is shown to be effective in controlling the trade-off between the verification errors. For evaluation of proposed architecture, the ideal fusion error rates are calculated in chapter 4 using the expressions developed for FRR and FAR. These equations are derived under the assumption of statistical independence between the classifier decisions and so there exist a difference between the theoretical (ideal) and experimental error rates. This difference, in general, is because of the statistical dependence between the classifier decisions. The dependence in this dissertation is modelled using correlation between the classifier decisions.

The exact class-conditional error rates for the fusion of correlated decisions were estimated using the full expansion of Bahadur-Lazarsfeld Expansion (BLE). The expressions for the error rates of the multi-instance fusion and multi-sample fusion schemes are modified

to incorporate the correlation between the classifier decisions. The error rates for multi-instance fusion were developed considering the conditions of acceptance from each of the '*n*' decisions. Similarly, the multi-sample fusion error rates were expressed using the BL expansion and the vector of rejections from multiple samples. The expressions for multi-instance and multi-sample fusion schemes were integrated for determining the proposed fusion verification error rates, i.e., the multi-sample fusion error rates were substituted as base errors in the expressions for multi-instance fusion.

The dependence is '*favourable*' when error rates after fusion using either *'AND'* or *'OR'* rules were smaller than fusion of independent classifier decisions. The *ideal false rejection rates* (FRR) were observed to be higher than *experimental/predicted FRRs* whereas ideal *false acceptance rates* (FAR) were lower than *experimental/predicted FARs* for multi-instance fusion. The results are complementary for multi-sample fusion. The difference between the ideal and predicted error rates for proposed architecture decreases with an increase in instances and samples used for fusion as the correlation progressively reaches zero.

As the correlation coefficient for an '*n*'th instance is dependent on the previous on 2^{nd}, 3^{rd} ..., $(n-1)$th order coefficients, the relationship with error difference (between ideal & predicted error rates) is not direct. The complete theoretical and experimental analysis to identify the conditions for favourable dependence of '*n*' correlated classifier decisions was presented in this chapter. When the 2nd-nth order impostor correlation coefficients are of same sign and negative, the dependence is favourable whereas positive 2nd-nth correlation coefficients are favourable on client decisions. For correlation coefficients with different signs, favourable dependence is determined using correlation coefficients between instances, correlation between samples for each instance and base error rates for instances.

The multi-instance fusion performance is better when impostor and client-impostor favourable digit combinations are considered as FAR decreases with an increase in digits. On the other hand, client and client-impostor favourable digits have shown improved performance as FRR decreases with multiple samples. With favourable client and impostor combinations, the experimental error rates were always lower than ideal error rates. Nevertheless, the classifiers in these favourable combinations do not ensure optimal fusion performance. The next chapter explains the methods in which the best set of classifiers/instances can be selected for better fusion performance.

Chapter 6
Classifier selection for the proposed fusion using '*Sequential Error Ratio'* criterion

6.1 Introduction

In the previous chapter, the statistical dependences between classifier decisions that are favourable for the proposed sequential fusion scheme were investigated. The use of favourable digit combinations for speaker verification was empirically shown to result in better performance than theoretically obtained errors under independence assumption. However, fusion of these favourably dependent decisions may not ensure optimal or best possible performance. This chapter thus deals with a method to select the classifiers that when combined maximizes the accuracy/performance of proposed fusion design.

The fusion of multiple classifiers (or decisions) has been shown to be an effective solution for difficult pattern recognition tasks. When a divide-and-conquer approach is used for combining multiple classifiers, the new inputs are directed to a specific classifier that performs well for each input type. In a sequential fusion scheme, a single best classifier is initially used and other classifiers are involved only if it fails to provide a reliable decision with sufficient confidence. The selection of the best among these schemes depends on application requirements and thus the design of a fusion scheme emphasizes selection of the best classifier for a given application. The design is then supplemented when different classifiers are invoked for different inputs with consideration for factors such as user-dependent and class-dependent information, cost of application and specific knowledge about the input.

An efficient fusion design, in general, requires the optimization of combination method (*'decision optimization'*) and then selection of an optimal set of classifiers (*'coverage optimization'*) [25]. The choice of combination methods depends on factors such as type of classifier outputs, number of classes and amount of training data. Further, for a large pool of different classifiers, there are a number of possible combination strategies. The combination method, sequential decision fusion using '*AND & OR Rules*', has been evaluated for set of classifiers in previous chapters (chapter 4 & 5). As combining all classifiers has been shown

in literature to be expensive and rarely optimal, various classifier selection methods that are employed for evaluation of improvement in fusion performance are investigated in this chapter.

As single best performing classifier does not guarantee the optimal or the best possible performance, the design of fusion system requires selection of a subset of classifiers that produce an optimal possible performance for a particular combination method. Two types of classifier selection techniques proposed in the literature are explained in section 6.2. In static classifier selection, the best subset of classifiers is found prior to classifying any test pattern [217]. The other criterion is dynamic selection where the classifier subset is dependent on test pattern being classified [218]. In Section 6.3, the accuracy and diversity related evaluation criteria used for dynamic selection of classifiers for optimal performance are investigated and evaluated for text-dependent speaker verification. A new criterion is proposed, in section 6.4, for the selection of classifiers at each stage of fusion. The proposed measure is then evaluated for the variation in fusion performance of classifiers with similar and different performances. Final section 6.5 presents the experimental results of the fusion of classifiers with and without repeated samples, selected using the evaluation criteria.

6.2 Classifier Selection Methods

Hybrid fusion systems based on the combination of multiple classifiers are used to improve performance in high pattern-recognition applications [219]. Optimal performance for the design of fusion architecture is based on the combination method and the selection of a classifier subset [25]. One method to achieve optimization is to design the architecture such that optimal classifiers are carefully modelled and then an optimal combination method for the classifier is determined. The choice of combination methods that can be used here is large ranging from simple voting rules through to trainable combination functions [219]. Another design method is to decide on a fixed, simple decision combination function and then select mutually complementary/diverse classifiers that when combined can achieve optimal performance [187].

In this dissertation, the combination method is fixed (i.e., sequential fusion using *'AND and OR Rules'*) and a classifier is selected at each stage of fusion that results in best possible performance. One possible training approach for classifier selection is to train all classifiers on whole training data set and then select a best classifier for each pre-specified

region [220]. Another method, used in this dissertation, is to specify the region of input data first and then train a responsible classifier for each region [221]. As details regarding the fixed combination method are covered in chapters 4 and 5, the various methods in which classifiers are selected for optimal performance are discussed here.

The assumption about classifier selection is that each classifier is *an expert* on some part of input data and is, therefore, responsible for given pattern during classification. As training and classification methods used for verification are predefined, the selection here concerns to determination of the best set of classifier models trained on different data. In the classifier selection literature, two types of classifier selection systems have been distinguished: Static and Dynamic [217].

6.2.1 Static Classifier Selection

Static classifier selection has been discussed in several studies [217, 222]. In this method, the input data selected is specified during training or evaluation stage, prior to classification of unseen data. For classifier selection, two strategies can be utilized. In first scheme, the appropriate classifiers can be assigned for pre-partitioned sections of data using clustering [222]. Then the classifier with highest estimated accuracy is selected for each input data space. Second approach is based on the classifier, i.e., a region is found where each classifier has its best performance [217]. The *Evaluation and Selection* approach has been commonly used for static classifier selection method. The selection is initially determined and evaluated on tune/validation/development set and is fixed for classification of unseen data. As the static selection of classifiers is based on average performances of known data in development set, there is always the possibility that same selection may not be well adapted for test set with unknown data.

6.2.2 Dynamic Classifier Selection

Most combination methods for multiple classifiers are based on the assumption of independent errors from different classifiers. As the design of such a classifier set is difficult in real pattern recognition applications [219], the dynamic classifier selection (DCS) approach is proposed to avoid the error independence assumption [223]. Srihari et al. [4] introduced the concept of dynamic classifier selection as an alternative to combination in multiple classifier systems. In this approach, classifiers are selected during classification based on the training performances and the parameters of data to be classified. For each

pattern/data sample, classifier that most likely classifies the pattern correctly is selected. Woods et al. [39] proposed the selection of classifiers based on local accuracy estimates. They have shown that when individual classifiers are optimized, the overall performance can be improved significantly by the dynamic selection of classifiers using local accuracy. Giacinto and Roli [218] proposed the concept of adaptive selection of multiple classifiers in order to select the most appropriate classifier for each input pattern.

Kuncheva [224] devised a simple clustering-and-selection algorithm based on a probabilistic interpretation of classifier selection. In this approach, dataset is clustered and the most successful classifier for each cluster is selected. In [217], the combination of classifier selection and fusion approach is compared empirically against switching between selection and fusion with discussion on the differences in classifier set using static or dynamic selection method. Liu and Yuan [225] also proposed an algorithm where feature space is partitioned by clustering separately the correctly and the incorrectly classified samples for each classifier. The *'Cluster and Select'* approach is in between Static and Dynamic Classifier Selection. The dataset is divided into regions in advance during the training stage and thus considered static. However, the classifiers are selected dynamically depending on the most appropriate region for the new sample/pattern. Therefore, both static and dynamic selection approaches do not have to be exclusively applied for multiple classifier system. An approach similar to this is employed in this work where the dataset is divided into different digit regions and a separate HMM digit model is trained in advance. Nevertheless, the classifiers used for fusion are dynamically selected at each decision stage of the architecture. The next section presents the investigation on best criterion used for classifier (instance) selection in the multi-instance fusion scheme.

6.3 Classifier Selection Criterion

Ho [226] studied the complexity of classification and comparative advantages of different multiple classifier system designs. Although some design methods have proven to be very effective, clear guidelines for choosing the best design method for classification are not yet available. The designs are varied by coupling different techniques for selecting classifiers with different combination functions, but the best classifier selection/combination is determined by performance evaluation. The 'overproduce and choose' or 'test and select' approach has been proposed for the most appropriate multiple classifier/hybrid system design

[41]. The basic idea is to select the subset of classifiers, that when combined can achieve optimal accuracy, from an initial large set of 'candidate' classifiers (overproduce phase). The computational complexity of the choice phase can be limited by using the appropriate selection criterion [41, 227]. The most commonly used criteria for selection of a classifier subset that results in optimal performance are investigated in this section.

6.3.1 Heuristic Rules

Individual best performance has always been considered a universal indicator for selection of the best classifiers. This method is preferred in most industrial applications because of its simplicity, reliability and robustness [43]. The major problem with individual performance criteria is the inconsistency in evaluation. When more and more classifiers with low performance are included, the resulting selection can produce only worse combinations. So the individual classifier and combination function performances are evaluated for appropriate selection.

Partridge and Yates [41] proposed some techniques that exploit heuristic rules for choosing classifier sets based on classifier performance. A simple technique is the selection of '*k*' best classifiers, *'Choose the Best'*, where *k* classifiers with the highest classification accuracy are selected from set. The heuristic rules are mainly validated for the classifier subsets that exhibit similar degrees of error diversity. This method initially sorts the classifiers based on their performances and then determine the number of best-performing classifiers required for optimal performance. The selection requires the computations of a single best classifier, a pair of the best classifiers, the best three classifiers and so on up to '*n*' classifiers in set. The complexity of this rule is in linear order of $O(n)$, although for each combiner performance (2*n*-1) combinations are to be evaluated. Although the computational complexity of the classifier selection is greatly reduced using these rules, the optimality of such heuristics is considered to be far from being guaranteed [41].

Another approach is the use of combination method performance rather than individual classifier performances as the selection criterion. This method of selection is precise and meaningful with consistent comparisons of different classifier subsets regardless of the number of classifiers and their individual performances. This approach, however, considers the *'Evaluation and Selection'* method, as the combinations with optimal performance are chosen by the selection algorithm and is evaluated for either the training or

the validation/tune sets with known data samples. Therefore, there exists the possibility that selected combinations may not necessarily remain optimal for unseen data.

The selection of individual classifiers also presents with generalization problem resulting in performance degradation of selection algorithm. This decrease in performance is limited by estimating the selection algorithm on the data space, which has not been used for selection of either the optimal combination function or individual classifiers. As a result, the static selection of classifiers here requires larger than usual training/tune datasets for obtaining reasonable reliability. Further, the complexity of such performance driven selection can be of exponential order in the case of exhaustive evaluations for possible classifier subsets. If the combiner is itself complex here, it could drastically slow down the search algorithm or even lead to intractability for larger numbers of classifiers.

6.3.2 Sequential Search Algorithms

As combining a pair of the best classifiers may not always be the optimal selection, greedy approaches/search algorithms have been used in literature that concentrates on adding or removing a specific classifier until maximum combiner performance is achieved. In this approach, an objective/evaluation function - i.e., individual classifier performances or diversity measures, is used for selecting classifiers that optimize the performance [228]. Among the existing search algorithms, the two most commonly used include sequential forward and sequential backward search algorithms [229]. Sharkey et al. [227] proposed an exhaustive search algorithm with the assumption of having a small pool of classifiers, whereas Ruta and Gabrys [43] proposed the use of sequential search algorithms for static classifier ensemble selection.

6.3.2.1 Sequential Forward Search

The sequential forward search starts from a single classifier that can be either randomly selected or the classifier with best accuracy. The next classifiers are iteratively selected based on the combination with lowest/highest value of the evaluation function. The optimization process is stopped when there is no improvement on improving the fusion performance or if all the classifiers are selected. In static selection, the selected classifier set at end of search for validation dataset is used to classify the test dataset.

6.3.2.2 Sequential Backward Search

The sequential backward search starts from the full classifier set. A classifier is eliminated iteratively based on the evaluation function. The classifier that makes smallest contribution to the reduction of error, is eliminated. The optimization process is stopped, if all the remaining classifiers are tested to be significant or only one classifier remains. Similar to forward selection, the classifier selection set at end of search for tune dataset is used to classify the test dataset. The evaluation of selection criterion for all possible pairs of remaining classifiers imposes quadratic complexity. The overall complexity of these selection algorithms is, therefore, of the order $O(n^3)$. Despite its relatively high complexity, the search algorithms do not guarantee the optimality of combination found.

6.3.3 AdaBoost

Viola and Jones [5] pointed out that classifier learning is analogous to classifier selection and many instances of classifier training may be replaced with optimal classifier selection using methods such as boosting or bagging. Magee [230] proposed the use of AdaBoost method where classifiers are ranked in order with an associated weight. This method used the learned statistics from input data for selection of an ‘optimal’ subset of classifiers from a finite set by picking the first '*n*' or putting a lower limit on weight. The AdaBoost algorithm [231] is therefore used for multiple classifier training and multiple classifier selection and combination [230, 232]. The algorithm used in this dissertation for selection of classifiers using AdaBoost ([232]) is given in table 6.1. The general termination condition for selection is based on the combination error $(\varepsilon \leq 0.5)$. However, for multi-instance fusion, the errors (false rejection rate) soon increase and the error condition may not be satisfied for most of the combinations. Therefore, the selection here is terminated only when the number of classifiers selected is equal to classifiers required for fusion.

The algorithm used for classifier selection is applicable for two-class classification where the definition of *'correct classification'* implies both true acceptances (true positives for client decisions) and true rejections (true negative for impostor decisions) whereas false rejections (false negative for client decisions) and false acceptances (false positives for impostor decisions are considered *'incorrect classifications'*. The selection of classifiers here can be based on the importance of true acceptances, true rejections or the combination of

Table 6.1 AdaBoost algorithm for classifier selection based on minimum weighted errors

- Consider a set of labelled decisions for input $(x_1, y_1), (x_2, y_2), ..(x_n, y_n)$, where $y_i = 0,1$ for impostor and client decisions respectively.
- Initialise equal weights for all the client and impostor decisions.
- Repeat (until '*n*' classifiers are selected from '*s*' classifiers; $n \leq s$):

  1. Normalise the weights $w_i \leftarrow \frac{w_i}{\sum_{j=1}^{s} w_j}$

  2. Select the classifier C_j with the lowest error $\varepsilon = \sum_{i=1}^{s} w_i \, | C_j(x_i) - y_i |$

  3. Update the weights

  $$w_{i_{new}} = w_i \beta_t^{1-e_i}$$

  where $\beta_t = \frac{\varepsilon}{1-\varepsilon}$ and $e_i = \begin{cases} 0, & y_i = 1 \\ 1, & y_i = 0 \end{cases}$

both. As weights are used for classifier selection, the set of classifiers that result in the best possible performance with a minimum number of computations are selected. When multiple samples are allowed for fusion of multiple instances, the instances selected at each decision stage tends towards the ones that require less number of repeated samples.

6.3.4 Diversity Measures

Instead of selecting classifiers based only on their accuracy/classifier performance, the diversity between classifiers is employed as an evaluation criterion for selection. The expressions for the most commonly used measures are listed in table 6.2 that are grouped into a pair wise and non-pair wise diversity measures [42]. The pair-wise measures are based on the measurement of diversity between any pair-wise classifiers, e.g. Q-statistics, kappa statistics, correlation coefficient, disagreement measure and double-fault measure. For classifiers more than two, the averaged values are calculated (table 6.2). The non-pair wise diversity measures are calculated simultaneously for '*n*' classifiers (n>2), e.g., entropy, kohavi-wolpert variance, measure of difficulty, generalized diversity and coincident failure diversity.

Table 6.2 Expressions for pairwise and non-pairwise diversity measures [42] for dynamic classifier selection

Name	Expression
Q-Statistic	$Q_{avg} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q_{i,j};\ Q_{i,j} = \frac{N^{11}N^{00} - N^{10}N^{01}}{N^{11}N^{00} + N^{10}N^{01}}, i \neq j$
Correlation Coefficient	$p_{avg} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_{i,j};$ $p_{i,j} = \frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{(N^{11}+N^{10})(N^{01}+N^{00})(N^{11}+N^{01})(N^{10}+N^{00})}}, i \neq j$
Disagreement Measure	$dis_{avg} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} dis_{i,j};\ dis_{i,j} = (N^{10}+N^{01}), i \neq j$
Double Fault Measure	$DF_{avg} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} DF_{i,j};\ DF_{i,j} = N^{00}, i \neq j$
Kohavi-Wolpert Variance	$KW = \frac{1}{nL^2} \sum_{j=1}^{n} \sum_{i=1}^{n} d_{i,j} \left(n - \sum_{i=1}^{n} d_{i,j} \right)$
Interrater Agreement	$k = 1 - \frac{\sum_{i=1}^{N} \left(n - \sum_{i=1}^{n} d_{i,j} \right) \sum_{i=1}^{n} d_{i,j}}{mn(n-1)P(1-P)}$
Entropy Measure	$E = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{\lceil n/2 \rceil} \min \left\{ \sum_{i=1}^{n} d_{i,j},\ \left(n - \sum_{i=1}^{n} d_{i,j} \right) \right\}$
Measure of Difficulty	$k = 1 - \frac{\sum_{i=1}^{n} \left(\left(\sum_{i=1}^{n} d_{i,j} \right) \left(n - \sum_{i=1}^{n} d_{i,j} \right) \right)}{n(n-1)\overline{p}(1-\overline{p})}$
Generalised Diversity	$GD = 1 - \frac{p(2)}{p(1)}$ where $p(1) = \sum_{i=1}^{n} \frac{i}{n} p_i$ & $p(2) = \sum_{i=1}^{n} \frac{i}{n} \frac{(i-1)}{(n-1)} p_i$
Coincident failure diversity	$CFD = \begin{cases} 0, & p_0 = 1 \\ \frac{1}{1-p_0} \sum_{i=1}^{n} \frac{n-i}{n-1} p_i, & p_0 < 1 \end{cases}$

n: total number of base classifiers; m: number of input samples
P: average accuracy of the base classifiers; p_i: accuracy of the base classifier c_i
$N^{ij}: i-$decision for classifier C_i, $j-$decision for classifier C_j

Different measures of diversity have been shown to select the best set of classifiers [42, 233, 234]. Kuncheva [235] presented a review on works where diversity has been utilized to select the final set of classifiers. Goebel and Yan [236] proposed the use of correlation measures for optimal classifier selection from a given set of classifiers to reduce the amount of exhaustive evaluations. Giacinto and Roli [237] formed a pair-wise diversity matrix using the *double fault* measure and the *Q statistic* [238] to select classifiers that are least related. Ruta and Gabrys [234] have shown that diversity measures for majority voting are particularly good at reducing of system complexity but imprecise and limited to the lower order dependencies. Moreover, measuring the diversity of member classifiers is by no means trivial, and there is a trade-off between diversity and member accuracy. The experimental evidence presented in literature [42, 234, 239] have shown that there exists a weak correlation between diversity measures and combined performance. The major risk of using them as selection criteria is simply picking the most diverse and not best-performing combinations. Another issue arises with possibility that the diversity measures of classifiers can be altered based on the combination methods. Tumer and Ghosh [87, 210] have shown that under certain assumptions, the averaging combination method produces accuracy which is related to the correlation between classifier outputs. They extended this result to show similar relationship for combination by order statistics of minimum, maximum and mean [73]. Shipp and Kuncheva examined the relationship between several widely used combination methods and several diversity measures [239]. However, there is no theoretical proof of any relationship in the general case. Further, the correlation between these measures of diversity and combination methods is not very high or consistent and thus the question of participation of diversity measures in designing classifier ensembles is still open.

Shipp and Kuncheva [239] have also concluded that directly calculating the accuracy for chosen combination method makes more sense than calculating the diversity and trying to predict accuracy. Even if the measure of diversity is easier to calculate than some combination methods, the ambiguous relationship between diversity and accuracy discourages optimising the diversity. One avenue that might suggest a useful method for

building classifier teams based on diversity is finding a more precise formulation of the notion of diversity and thereby constructing a more practical measure.

6.3.5 Experimental Results

The classifier selection methods are evaluated on datasets of *SET-1*, *SET-2* and *SET-3*. The protocol used for evaluation of these sets is described in section 3.5.2. The evaluations in this chapter are presented using the pooled results for each dataset of a speaker, if otherwise specified, is the pooled results for entire *SET*. In this section, the term '*classifier'* selection refers to the selection of an appropriate *'digit'* modelled using a HMM or the selection of an *'instance'* in the proposed architecture. The selection techniques are initially evaluated for fusion of multiple instances without repeated samples. The fusion performance is represented using the total error rate (TER) measure, however, for the ***multi-instance fusion*** the total error rates (false rejection rate) increases and ***soon exceeds 50%***. This increase in error rates is better controlled when multiple samples are allowed at each stage of multi-instance fusion. Therefore, selection here is continued until the number of classifiers selected is equal to the classifiers required for fusion. The objective is to determine the best criterion for selection of classifiers with *optimal performance*, even though the *best possible error rates* for combination is greater than 50%.

- ***Heuristic Techniques***

Table 6.3 presents the error rates for multi-instance fusion based on Choose '*k*' Best Rule for the test datasets of *SET-1*. A classifier at each instance or decision stage is selected using base performances, i.e., the pooled total error rates across the *SET*. As with the case of multi-instance fusion, the fusion of multiple decisions using *'AND Rule'* reduces false accepts but increases false rejects. The TER (FRR +FAR) increases with digits used for fusion as the increase in false rejects are greater than the decrease in false accepts. The single best classifier, '*k*' = 1, is selected by choosing the digit with minimum TER in the dataset. Suppose that the sequential approach combines two classifier decisions, the error probability is lower for the fusion of only the best classifiers selected rather than two randomly selected classifiers using 'Choose *k* Best' (*k*>2). For example, in table 6.3, fusion TER is observed to be lower for situations where only the best-chosen classifiers are combined, i.e., cases where the number of the best classifiers chosen are equal to the number of fused classifiers (TERs

across the diagonal in table 6.3). Although this method of selection is simple, the fusion of selected classifiers may not always result in the best possible performance.

An alternative to 'Choose the Best' is the '*best combination performance'* rule that enables selection of classifiers with the optimal or best possible fusion performance. Table 6.3 also presents the TER for use of the *best combination performance* as heuristic rule for classifier selection. At each decision stage, the *best combination performance* is selected for a classifier combination with lowest error rates from the entire set of possible combinations. For each selection of a classifier from a set of *'n'*, $(2^n - 1)$ comparisons are required - e.g., the selection of three-digit combination with lowest error rates requires almost 126 comparisons. Although the performance obtained using *'best combination performance'* is optimal, this rule requires exhaustive evaluations with increases in the search complexity for larger classifier

Table 6.3 Total Error Rates for the fusion of *'n'* digits selected using the heuristic rules - *'Choose k Best'* and *'Best Combination Performance'*

Heuristic Rule		Total Error Rate (TER) for fusion of '*n*' Digits					
		'*n*' = 1	'*n*' = 2	'*n*' = 3	'*n*' = 4	'*n*' = 5	'*n*' = 6
Choose '*k*' Best	'*k*' = 1	$0.258^{\pm 0.2}$					
	'*k*' = 2	$0.300^{\pm 0.2}$	$0.294^{\pm 0.2}$				
	'*k*' = 3	$0.340^{\pm 0.3}$	$0.330^{\pm 0.2}$	$0.372^{\pm 0.2}$			
	'*k*' = 4	$0.370^{\pm 0.2}$	$0.360^{\pm 0.2}$	$0.401^{\pm 0.2}$	$0.447^{\pm 0.2}$		
	'*k*' = 5	$0.400^{\pm 0.3}$	$0.388^{\pm 0.2}$	$0.429^{\pm 0.2}$	$0.474^{\pm 0.2}$	$0.515^{\pm 0.2}$	
	'*k*' = 6	$0.438^{\pm 0.3}$	$0.424^{\pm 0.2}$	$0.468^{\pm 0.2}$	$0.516^{\pm 0.2}$	$0.561^{\pm 0.2}$	$0.604^{\pm 0.2}$
	'*k*' = 7	$0.476^{\pm 0.3}$	$0.461^{\pm 0.3}$	$0.507^{\pm 0.2}$	$0.557^{\pm 0.2}$	$0.602^{\pm 0.2}$	$0.642^{\pm 0.2}$
Best Combination Performance		$0.258^{\pm 0.2}$	$0.286^{\pm 0.2}$	$0.356^{\pm 0.2}$	$0.429^{\pm 0.2}$	$0.503^{\pm 0.2}$	$0.586^{\pm 0.2}$

Table 6.4 Total Error Rates for the fusion of digits selected using *'choose k best'* (*k*=*n*) and *'Best Combination Performance'* rules for datasets of *SET-2* and *SET-3*

Number of digits ('*n*')	***SET-2***		***SET-3***	
	Choose '*k*' Best ('*k*' = '*n*')	Best Combination Performance	Choose '*k*' Best ('*k*' = '*n*')	Best Combination Performance
'*n*' = 1	$0.302^{\pm 0.20}$	$0.302^{\pm 0.20}$	$0.391^{\pm 0.10}$	$0.391^{\pm 0.10}$
'*n*' = 2	$0.350^{\pm 0.12}$	$0.342^{\pm 0.19}$	$0.408^{\pm 0.10}$	$0.396^{\pm 0.10}$
'*n*' = 3	$0.382^{\pm 0.19}$	$0.372^{\pm 0.18}$	$0.432^{\pm 0.10}$	$0.422^{\pm 0.10}$
'*n*' = 4	$0.420^{\pm 0.18}$	$0.403^{\pm 0.17}$	$0.460^{\pm 0.12}$	$0.446^{\pm 0.10}$
'*n*' = 5	$0.452^{\pm 0.16}$	$0.434^{\pm 0.16}$	$0.491^{\pm 0.11}$	$0.470^{\pm 0.10}$
'*n*' = 6	$0.481^{\pm 0.14}$	$0.469^{\pm 0.15}$	$0.515^{\pm 0.10}$	$0.496^{\pm 0.10}$
'*n*' = 7	$0.523^{\pm 0.15}$	$0.510^{\pm 0.15}$	$0.541^{\pm 0.11}$	$0.522^{\pm 0.10}$
'*n*' = 8	$0.560^{\pm 0.12}$	$0.547^{\pm 0.12}$	$0.574^{\pm 0.12}$	$0.552^{\pm 0.12}$

pool. Therefore, error rates for the *'best combination performance'* rule are lower compared to that of '*n*' classifier fusion selected using 'Choose *k* Best' where *k*=*n*. The above conclusions of classifier selection using heuristic rules are extended to evaluations on other datasets of *SET-2 & SET-3*. Table 6.4 presents the total error rate for the fusion of '*n*' digits selected using the heuristic rules - choose *'k'* best (*k*=*n*) and best combination performance of *SET-2 & SET-3*. The total error rates for *'best combination performance'* rule is optimal in the sense that the particular combination results in the lowest possible error rates for speaker verification performed on test datasets. Therefore, these total error rates are considered as reference for performance evaluations of selection criteria for multi-instance and multi-sample fusion scheme.

➢ *Sequential Search Algorithms*

The static classifier selection approaches based on individual performances does not consider the effect of individual classifiers on the combination performance. If selection is based on just the best combination performance, the complexity of search increases exponentially. An alternative is to use both the best classifier performances and the best combination performance for selection. One such approach is to initially select the best

classifier based on performance and then dynamically select classifiers at each stage based on the performance of the combination (sequential forward selection). Another method is to find the best combination and then dynamically eliminate classifier that results in a high number of errors at each stage (sequential backward selection).

Figure 6.1(a), (b) and (c) presents the fusion performance (total error rates) for

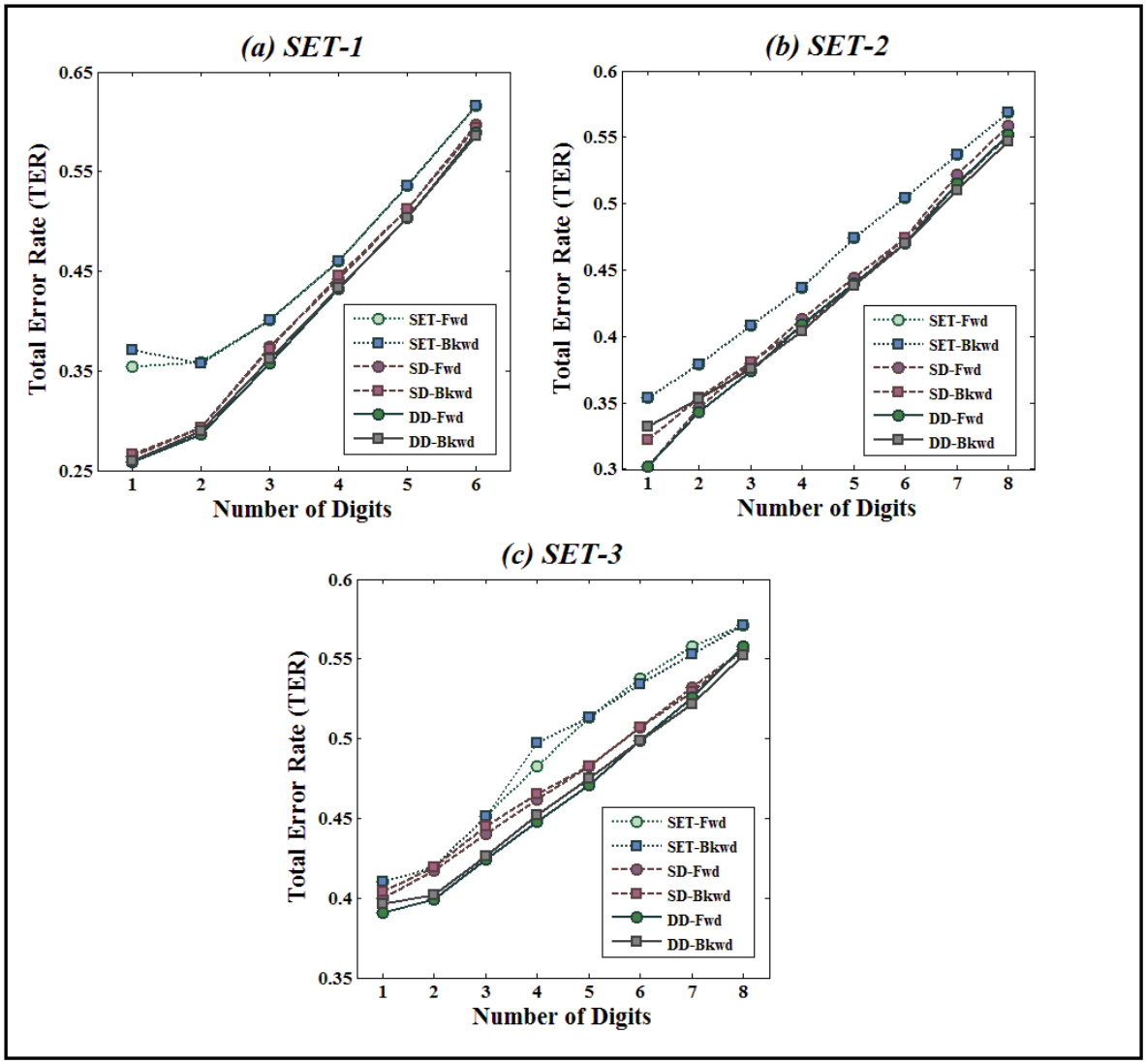

Figure 6.1 Total error rates for digit selection using heuristic sequential forward and backward search algorithms (a) *SET-1*, (b) *SET-2* and (c) *SET-3*.(SET-Error rates pooled for all speakers in a set, SD (speaker-dependent)-error rate pooled for each speaker from all datasets, DD (dataset-dependent)-error rates for each speaker from individual datasets; Fwd - Sequential Forward Search, Bkwd - Sequential Backward Search)

classifier selection using the sequential forward and backward selection algorithms for test datasets of *SET-1*, *SET-2* and *SET-3* respectively. The classifier with the best base performance (lowest error rates) is used as starting point for forward search algorithm. The next classifier at each stage is determined using the *best combination performance* as evaluation function. The digit combination with best performance (lowest error rates) is used as the starting point for backward search algorithm. The classifier that results in highest error rates for the digit combination is eliminated at each stage. The total error rates are presented for three types of dependencies - pooled results for all the speakers in a set (SET), speaker-dependent (SD) and dataset-dependent (DD). In case of SET dependence, the next digit/classifier selected is the same for all client and impostor speaker samples. For the speaker dependence (SD) case, the classifier selected for next stage is the same for each speaker but may be different among speakers. In dataset-dependence (DD), the classifiers are selected for each speaker dataset in a *SET* and so the digits selected at each stage can be different for each dataset of same speaker.

The total error rates for forward and backward search algorithms are lower for dataset-dependent and speaker-dependent classifier selection rather than selecting a classifier at SET level. Although dataset-dependence results in the best performance, it is difficult to generalize the selected classifiers at each stage. Therefore, speaker dependent selection is considered here as an alternative for achieving better performance at each stage with the same classifiers selected for each speaker. The total error rates (TERs) for both forward and backward selection algorithms are observed to be similar for test datasets of *SET-1*, *SET-2* and *SET-3* (fig. 6.1). One reason for this might be that limited number classifiers are available in the datasets. As the performances are similar for both sequential algorithms, the sequential forward selection algorithm is used for further performance evaluations of multi-instance fusion. The forward selection approach is also better applicable for proposed architecture where the number of classifiers/digits is dynamically selected for verification.

- ***Dynamic Classifier Selection***

The other selection criterion commonly used for dynamic selection of classifiers is the diversity between classifiers. The forward search algorithm with highest diversity as evaluation criterion is used for classifier selection. For these evaluations, digit with the best base performance is selected as first classifier in selection set. At each stage, the selection set is added with another digit - from the remaining classifier set that is highly

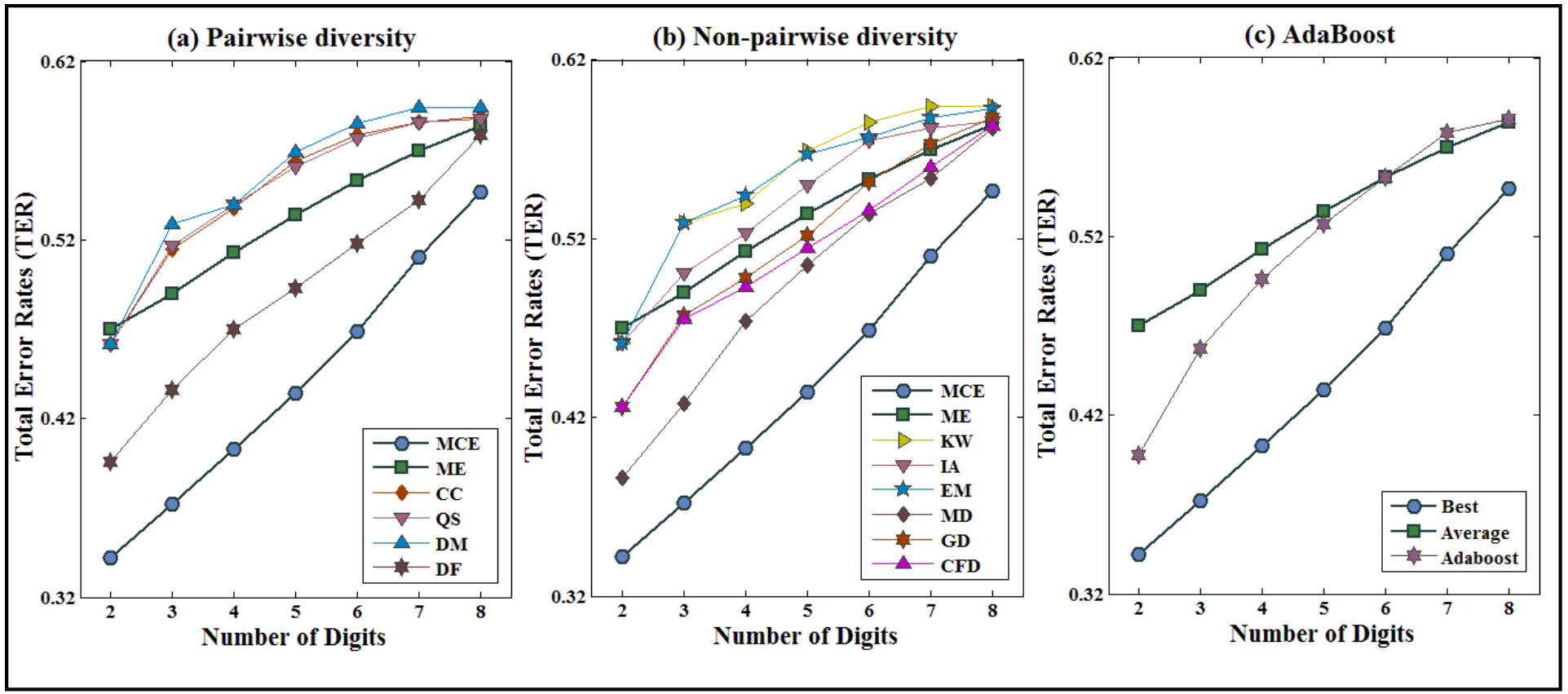

Figure 6.2 Total Error Rates for multi-instance fusion of classifiers selected using (a) Pairwise and (b) Non-Pairwise Diversity (c) AdaBoost (minimum weighted errors) measures as evaluation criteria for datasets of *SET-2* (MCE-Minimum Combination Error, ME-Mean Error, CC-Correlation Coefficients, QS-Q Statistic, DM-Disagreement Measure, DF-Double Fault, KW- Kohavi-Wolpert Variance, IA-Interrater Agreement, EM-Entropy Measure, MD-Measure of Difficulty, GD-Generalized Diversity and CFD-Coincident Failure Diversity)

diverse (or has minimum weighted error) with the prior selected digits. The selection process is terminated when all the available classifiers in the set are selected. The best performance for multi-instance fusion is obtained when the minimum combination error (MCE) is used for digit selection, whereas the measure Mean Error (ME) allows selection of digits that results in minimum mean error rates for the selected digit combinations [43]. The performance for fusion of digits selected using diversity measures [42] and AdaBoost are compared to that of MCE and ME. Figure 6.2(a) and (b) shows the TER for fusion of digits selected using pairwise and non-pairwise diversity measures as evaluation criteria whereas figure 6.2(c) presents the performance results for selection based on AdaBoost (minimum weighted error) for test datasets of *SET-1*. As demonstrated in [240], classifiers selected using the measures *double fault* and *difficulty* are shown to be better than Mean Error (ME) and are the best among the other diversity measures at predicting the best set of classifiers. The TER for combination of digits selected using MCE are observed to be equal to error rates obtained using the heuristic rule - *best combination performance* (table 6.3).

The double fault, difficulty and adaboost measures are shown to be reasonably good evaluation criteria but the differences in fusion performance compared to Minimum Combination Errors (MCE) are high. Further, this difference is observed to be inconsistent with an increase in digits used for fusion and the corresponding error rates soon reaches to that of ME measure. Although results presented in the figure are only for *SET-2*, the conclusions are extended to evaluations performed on the datasets of *SET-1* & *SET-3* (Figure 6.4). It is thus demonstrated that the existing diversity measures and adaboost weighting criterion for multi-instance fusion do not result in the best possible performance for each digit combination. Therefore, a new selection criterion is proposed in the next section that is specifically tuned to the characteristics of sequential fusion using *'AND & OR'* decision rules.

6.4 Sequential Error Ratio

The diversity measures discussed above can be basically defined using one of the three different approaches as explained in [241]. The measures used as selection criteria can be solemnly based on classifier outputs, irrespective of whether the related decisions are correct or incorrect. Another approach is based on the correct or incorrect decisions from classifiers, assuming the correct answers are known. Though the use of this approach is beneficial because of the use of correctness of knowledge, it neglects a large amount of

information by simply joining all incorrect classes. In the next approach, both the classification and correctness information is used for classifier selection. As the identical correct results are more desired than identical incorrect results, this method of selecting classifiers can be advantageous [242] .

The performance of sequential fusion scheme depends on two factors - the number of classifiers and the order in which classifiers are selected and combined for the best performance. To avoid disagreement between the design of combination method and selection approach, the selection criterion should be tuned to the combination design by exploring its characteristics. In this section, a new criterion is proposed for the sequential selection of instances and the integration of instances and samples.

For sequential fusion of instances proposed in [32, 36], the focal idea is that all classifiers agree on the correct decision (i.e., accept the claim) from a client. For an impostor, at least one decision should disagree with previous incorrect decisions (i.e., reject the claim). The total error rates for multi-instance fusion are better when the increase in false rejects is less than the decrease in false accepts. Therefore, the classifier with decisions that mostly agree rather than disagree with previous classifiers' correct decisions results in better overall performance. Based on these characteristics, a new measure called the sequential error ratio (*SER)* is proposed which is the ratio of the number of input samples on which classifier disagrees with previous correct decisions to the number of input samples on which classifier agrees with the previous correct decisions. The sequential error ratio for two classifiers C_i & C_j, is

$$SER_{i,j} = \frac{N^{10}}{N^{11}} \tag{6.1}$$

Where, $N^{ab}(\{a,b\} = \{0,1\})$ is the number of samples for which the decisions from C_i & C_j are '*a*' and '*b*' respectively**.** The output ***'1'*** here indicates that the classifier makes a ***correct decision*** whereas ***'0'*** indicates an ***incorrect decision*** from the classifier. The pair of classifiers with minimum SER is initially selected. Another approach is to select the first classifier, say C_i, based on maximum accuracy/minimum error. The classifier, say C_j, with *minimum SER* is selected at the second stage of fusion. The next classifier, C_k, in the sequence should also agree rather than disagree most of the time with decisions from previous classifiers, C_i & C_j.

The sequential error ratio calculated for three classifiers is independent of two-classifier measure and is given as:

$$SER_{i,j,k} = \frac{N^{110}}{N^{111}} \tag{6.2}$$

Similarly, at '*n'*th sequential stage, the classifier selected should agree most of the time, rather than disagree, with the previous '*n-1*' classifiers' correct decisions. The ratio used for the inclusion of nth classifier, C_n, in to the selection set $(C_1, C_2, C_3, ...C_{n-1})$ is defined as:

$$SER_{123...(n-1)n} = \frac{N^{111...10}}{N^{111...11}} \tag{6.3}$$

The classifiers selected using the information from client, impostor and client-impostor samples might be different. When the datasets used for evaluation are typically imbalanced, with one class underrepresented compared to other class with relatively large number of samples, the individual accuracy measure causes unreliable results [243]. However, the proposed *Sequential Error Ratio* has little relationship with individual classifier performances but exploits both the classification and decision correctness information of the selected classifiers. The performance of the sequential error ratio measure is evaluated for the multi-instance fusion with and without repetition of samples.

6.4.1 Multi-instance Fusion

In fig. 6.2, the measures double fault (DF), measure of difficulty (MD) and adaboost are shown to result in better performances next to minimum combination error (MCE) measure. Figure 6.3 presents the error rates for classifiers selected using the above measures and *Sequential Error Ratio* (*SER*) in comparison with Minimum Combination Error (MCE) and Mean Error (ME). Among the measures represented, the classifier selection based on *SER* is demonstrated to result in lower total error rates next to MCE for datasets of *SET-1* (fig. 6.3(a)), *SET-2* (fig. 6.3(b)) and *SET-3* (fig. 6.3(c)). The number of evaluations required for MCE is less for these *SETs* because of the limited number of available digits. However, for large classifier set, the computations required increases significantly and thus the use of *SER* for classifier selection improves fusion performance.

The error rates presented in fig. 6.3 are for the fusion of classifiers selected using measures that considers output information from both client and impostor data samples, i.e.,

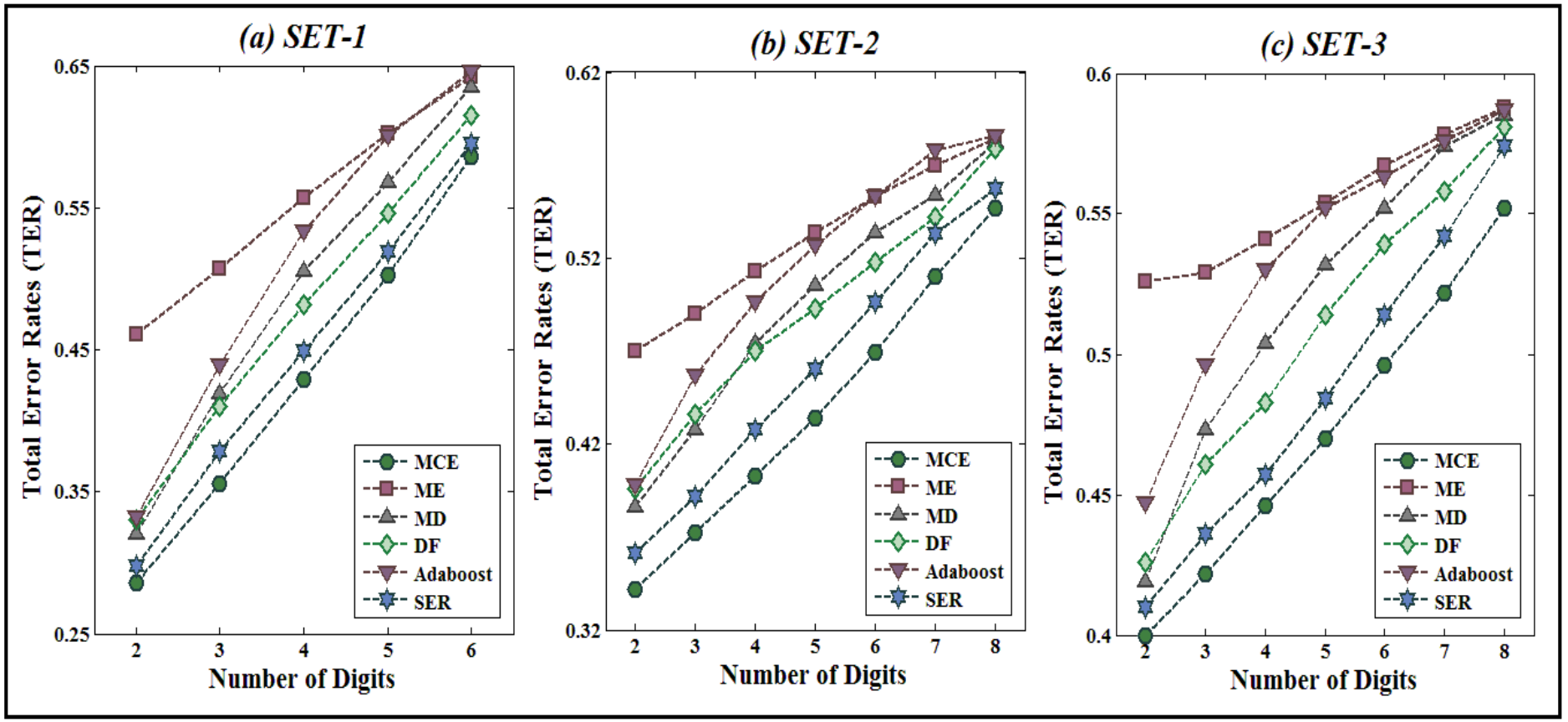

Figure 6.3 Total Error Rates for fusion of classifiers selected using diversity measures (DF, MD, MCE, ME), AdaBoost and SER for the datasets from (a) *SET-1*, (b) *SET-2* and (c) *SET-3* (MCE - *'Minimum Combination Error'*, ME - *'Mean Error'*, MD - *'Measure of Difficulty'*, DF - *'Double Fault'* and SER - *'Sequential Error Ratio'*)

the selection is based on a client-impostor dependent *SER* measure. When the selection is based on only the client or impostor output information, different set of classifiers with different fusion performances may be selected. The total error rates for multi-instance fusion of classifiers selected using client dependent, impostor dependent and client-impostor dependent measures are presented in fig. 6.4(a) whereas the effect of class-dependent *SER* on individual verification error rates (FRR vs. FAR) is represented in fig. 6.4(b). The client-dependent SER measure is observed to have better fusion performance compared to an impostor dependent SER and client-impostor dependent SER. The lower TER for client-dependent SER is due to the greater decrease in false accepts compared to an increase in false rejects (fig. 6.4(b)). Lower false rejects are obtained for the selection based on client outputs and lower number of false accepts are observed for an impostor dependent SER selection (fig. 6.4(b)). Although fig. 6.4 presents error rates for fusion of instances in SET-2, the conclusions are extended to other SETs with different datasets as well.

The error rates for sequential error ratio criterion are higher than the best possible combination performance mainly because the proposed measure considers only the decisions, which either agree or disagree with all the previous classifiers correct decisions. When one

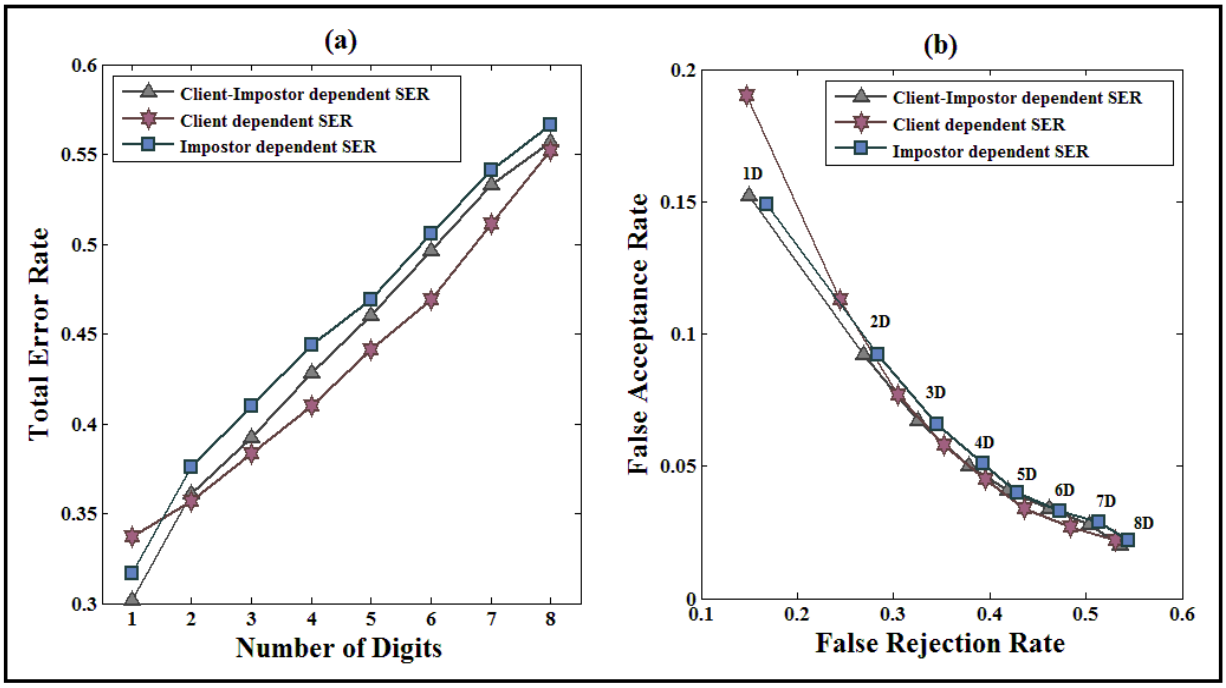

Figure 6.4 Error rates for multi-instance fusion of classifiers (digits-D) selected using client-dependent, impostor-dependent and client-impostor dependent sequential error ratio measure (a) Total Error Rate (b) False Rejection Rate vs. False Acceptance Rate

client decision from a set of selected classifiers is incorrect, then the combination, irrespective of the subsequent classifier decisions results in a false rejection. For impostor decisions, even when (at least) one or all of the previous classifier decisions are incorrect, the next classifier in sequence could make a correct decision thereby reducing the false acceptance rate. With the proposed SER measure, the decision combination situation where one or more previous classifiers are incorrect, but the correct decision is made by the classifier under consideration is not taken into account. For example, when three impostor decisions [111, 101, 110, 100, 001...] are combined, the error ratio does not take into account the samples with output sequences, such as 101 or 100, in which not all the previous classifiers make a correct decision. Therefore, the fusion of classifiers selected using SER may result in false accepts greater than the best possible performance for a combination, whereas false rejection rates are close to the best performance. Irrespective of this difference in false accepts, the SER measure is observed to be better than other classifier selection criteria. The total error rates (false rejection rate) for multi-instance fusion increases and soon exceeds 50%. This increase in error rates is better controlled when multiple samples are allowed at each stage of multi-instance fusion.

6.4.2 Multi-instance and Multi-sample fusion scheme

The proposed Sequential Error Ratio is demonstrated to be a better measure for dynamic selection of a digit at each stage in multi-instance fusion architecture. Although, the false acceptance rate decreases for fusion of multiple instances, the false rejection rate increases [32, 36]. The trade-off between FRR and FAR is controlled by integrating the multi-instance *('AND Rule')* and multi-sample *('OR Rule')* fusion schemes. In table 6.5, the total error rates are shown for combinations with *'n'* instances and *'m'* samples selected using Mean Error, Double Fault, Measure of Difficulty, AdaBoost and *SER* measures. The difference in error rates, for these measures compared to MCE, decreases with an increase in samples used for fusion. However, the SER selection results in lowest error percentages compared to other measures mainly because of the reduction in false accepts. Similar decrease in error rates are observed for double fault measure as well. The results presented in table 6.5 are for verification tests on *SET-2*but the conclusions are extended to other test datasets of *SET-1* and *SET-3 (*shown in fig. 6.5). The proposed measure is thus demonstrated to be an effective selection criterion to obtain better fusion performance irrespective of the number of classifiers or samples used for fusion.

Table 6.5 Error Rates for the MCE, SER, MD, DF and ME measures based instance selection and fusion approach with multiple samples for test datasets of *SET-2* (*n* - number of instances, *m*-number of samples)

Selection Criteria	Total Error Rate (in %) for (*n*, *m*)				
	(1, 1)	(2, 2)	(4, 2)	(6, 3)	(8, 4)
Minimum Combination Error	30.19	22.57	22.28	16.16	15.52
Sequential Error Ratio	30.19	23.49	24.56	16.81	15.59
Measure of Difficulty	30.19	27.16	28.14	19.60	16.54
Double Fault	30.19	23.72	24.99	17.51	16.20
AdaBoost	30.47	23.76	25.51	17.43	16.32
Mean Error	47.27	34.79	31.18	21.45	17.19

As with the case of favourable digit combinations, the digit sequence obtained using SER can also be speaker-dependent (speaker-specific). The sequence of the selected digits can be different between speakers as shown in table 6.6 but are similar for a speaker across different datasets. For example, the digit combinations in sequence 8-5-1-9-3-7-4-2, i.e., 85, 851,...85193742, selected using SER have better performance for Spkr-0047 whereas the selection of digits 3-1-7-6-9-5-8-2 each at a decision stage provides better performance for Spkr-0241in *SET-2*(test datasets). Further, the digit sequence with optimal performance for the same speaker may be slightly different in the order for the use of same measures (tune and test datasets of the same speaker in table 6.6). Therefore, the use of speaker-specific digit combination can be considered another measure to ensure reliable identity verification.

The evaluations presented above for dynamic classifier selection is based on the decision outputs for the known samples of a speaker. The best classifier set for an individual can be pre-determined on the known data (tune dataset) and later used in text-dependent speaker testing of unseen data (test dataset). The total error rates for fusion of digits selected using MCE and SER for evaluation and test datasets are represented in fig. 6.5. The error rates for tune dataset are obtained using the fusion of digits dynamically selected at each stage using MCE and SER measures. The error rates for test dataset, however, are obtained

for the digit sequence selected on tune dataset. Although the TERs for tune dataset are higher for the use of SER than MCE, the fusion performance for unseen test dataset is observed to be better for SER. In addition, the difference between tune and test datasets is higher for MCE measure than SER. Therefore, the proposed sequential error ratio measure is shown to be a better criterion for selection of classifiers with better performance for multi-instance and multi-sample fusion design.

The base error rates and digit sequence selected using SER on tune datasets can be used to predict the fusion performance of test dataset (unknown data). As the base classifiers for both tune and test datasets are in general assumed similar, the parameters (*n, m*) used to control the trade-off between FRR and FAR on tune dataset can be applied to test dataset. The fusion error rates for these two datasets can be expected to be slightly different because of the variations in decision correlation and the order of digit sequence. This difference in error rates can be used as another measure to determine the identity claim. If the difference in error rates for a particular digit combination is high, there is higher possibility that the unknown data being tested might not belong to the claimed speaker.

Although, better performance is shown for fusion of classifiers selected using SER measure, the actual fusion performance itself depends on classifier with similar or different performances and the dependence between the decisions [190]. The next section, thus, presents the empirical evaluation on performance of the proposed measure for selection of classifiers with similar and different performances.

Table 6.6 Optimal Digit Combinations for Tune and Test datasets of speakers from *SET-2*

Speakers	**Tune**	**Test**
Spkr_0047	8-5-1-9-3-2-7-4	8-5-1-9-3-7-4-2
Spkr_0074	8-1-3-7-5-2-9-6	8-3-7-1-5-2-9-6
Spkr_0086	3-5-1-2-7-9-8-6	3-5-1-8-9-7-2-6
Spkr_0176	3-1-5-7-8-2-9-6	3-1-7-5-8-2-9-6
Spkr_0241	3-1-7-6-9-5-8-2	3-1-7-9-8-6-5-4

6.4 Homogeneous Classifier Clusters

The accuracies of several competing classifiers are often compared to determine the most accurate classifier, or if all the classifiers are of similar accuracy then the classifier that is easiest to apply with less complexity can be selected. Looney [244] proposed a comparison strategy based on the application of repeated measures analysis techniques for dichotomous data whereas Goldstein [245] proposed the procedure applicable only when two classifiers are being compared using separate validation sets for each one.

The selection of a subset of classifiers that are significantly better than the rest (or than complete set of classifiers) is important for performance improvement. However, selection of the best cannot have a unique solution, as the notion of best can be quite subjective and depends on the criterion used. Though heuristic techniques have been used to order classifiers based on accuracy, the rules do not determine how many of classifiers are truly the best in the sense that they differ significantly from all the others. Tsoumada et al. [246] rephrased this problem of finding the best classifiers as the problem of finding a subset of classifiers with similar good performances with at least one classifier in the subset whose performance differs significantly from those not in the subset. This method enables to exclude classifiers with low performance, which may produce misleading results.

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or clusters. The cluster analysis methods[247] usually depend on fast and efficient combinatorial algorithms with the assumption of normality on test data. Scott and Knott [248] proposed a cluster analysis method for obtaining homogeneous groups of *n* classifiers based on the means of error rates. Gates and Bilbro [249] illustrated the use of Scott-Knott procedure in which the method first attempts to separate means into two groups. The two groups are then tested separately for additional separations, and the partitioning is continued until groups of single classifier or groups of homogeneous classifier (or both) are found. Tsoumakas et al. [246] also proposed the use of Scott and Knott procedure to determine the homogeneous group of classifiers with the smallest mean values of error rates. The Scott and Knott method used in this dissertation is similar to procedure explained in [246]. The procedure to determine the homogeneous group of classifiers with the lowest mean error rates is presented in Appendix A.3.

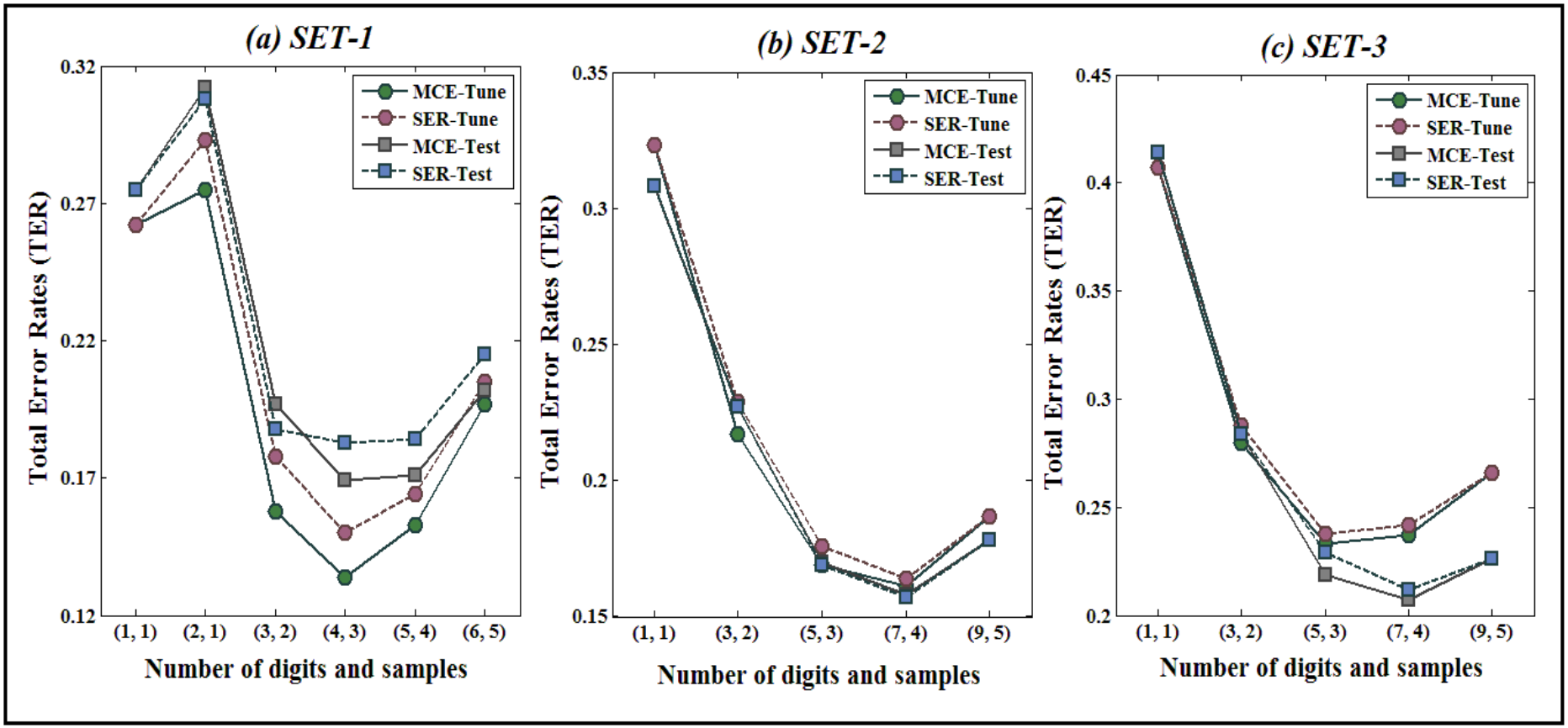

Figure 6.5 Total Error Rates for proposed fusion of classifiers selected using diversity measures (DF, MD, MCE, ME), AdaBoost and SER for the datasets from (a) *SET-1*, (b) *SET-2* and (c) *SET-3* (MCE - *'Minimum Combination Error'*, ME - *'Mean Error'*, MD - *'Measure of Difficulty'*, DF - *'Double Fault'* and SER - *'Sequential Error Ratio')*

6.4.3 Experimental Results

As demonstrated in previous sections, the use of accuracy, diversity measures or SER for classifier selection does not always ensure the best possible performance for *n*-digit combinations. One reason for this might be because of the inclusion of classifiers with low performance into fusion that may produce misleading results [246]. One approach to avoid the above situation is for classifiers in the dataset to be clustered or grouped, homogeneous with respect to their performance. The cluster selected by the Scott-Knott Clustering method should ensure that classifier performances are homogeneous with-in clusters and significantly different between clusters. For *SET-2*, the classifiers (3-1-8-5-7-2-9) in the cluster are found to be of significantly similar performances and at least one of the classifier in this cluster has significantly different performance compared to classifiers in a cluster (6-4). Although the *Cluster and Select* approach, in general, selects one classifier from each cluster, the approach used here considers te fusion of classifiers in the same cluster. The purpose of clustering classifiers here is to remove the classifiers with significantly low performance from the selection set to improve fusion performance. Therefore, the sequential forward selection algorithm is used for selection of classifiers within a cluster.

Table 6.7 presents the comparison of TERs for fusion of classifiers selected from the cluster and entire classifiers in dataset for *SET-2*. The fusion performance is observed to be better when classifiers are dynamically selected from clusters using the double fault measure and the measure of difficulty criteria. However, for selection based on SER, the total error rates are slightly higher when only classifiers with similar base performances are combined. For example, the fusion of five digits selected using SER results in 46% TER but the error increases to 46.8% when digits are selected only from the cluster. Even with the decrease in fusion error for classifiers (with similar performances) selected using the measures double fault and difficulty, lower TERs are obtained for fusion of classifier with SER criterion (with or without grouping the classifiers based on performances). The fusion of clustered classifiers neglects the complementary information from classifiers with significantly different performance. As shown in [190], the classifiers with very different performances can also result in improvement of fusion performance. Therefore, the proposed sequential error ratio measure is shown to be a better selection criterion for sequential fusion of classifiers with similar or different performances.

Table 6.7 Total Error Rates for Multi-Instance fusion of classifiers in cluster and all classifiers in the test dataset for *SET-2*

Number of Classifiers	Multi-Instance Fusion TER (in %)						
	MCE	Classifiers (1, 2, 3, 4, 5, 6, 7, 8, 9)			Classifier Cluster (3-1-8-5-7-2-9)		
		Difficulty	DF	SER	Difficulty	DF	SER
2	34.2	38.6	39.6	36.1	37.6	37.1	36.4
3	37.2	42.8	43.6	39.2	40.8	40.8	39.9
4	40.3	47.4	47.0	42.8	46.9	45.2	43.5
5	43.4	50.5	49.3	46.0	49.8	47.7	46.8
6	46.9	53.4	51.8	49.6	52.5	50.3	49.9
7	51.0	55.4	54.2	53.3	53.7	53.7	53.7

6.5 Error rates for *SER* selected digit combinations

The proposed architecture based on the sequential integration of multi-instance and multi-sample fusion schemes is empirically shown to improve the performance and allow a controlled trade-off between false alarms and false rejects. The fusion of *independent decisions* for all possible digit combinations and fusion of *dependent decisions* for favourable digit combinations were shown to reduce both the error rates simultaneously. However, the performance of fusion is improved when the best digit combinations are used at each stage (multi-instance fusion) of verification. Figure 6.5 presents the mean verification error rates (FRR and FAR) for digit combinations that are selected using client dependent, impostor dependent and client-impostor dependent SER measures for *SET-2*. The digit combinations selected using the sequential error ratio is shown to improve fusion performance. For example, the fusion of all possible three-digit combinations with three samples for each digit (5, 3) results in reducing the false rejection and false acceptance rate to 10.2% and 13.6% respectively. However, for digit combinations selected using *client-impostor dependent SER*, the FRR and FAR are reduced to 7% and 9.9% respectively. The FRR and FAR for parameter combinations (2, 2) selected using *client dependent SER* (in fig. 6.5 (c)) are 5.7% and 11.5% respectively whereas these errors are reduced to 8% and 10% for selection using *impostor*

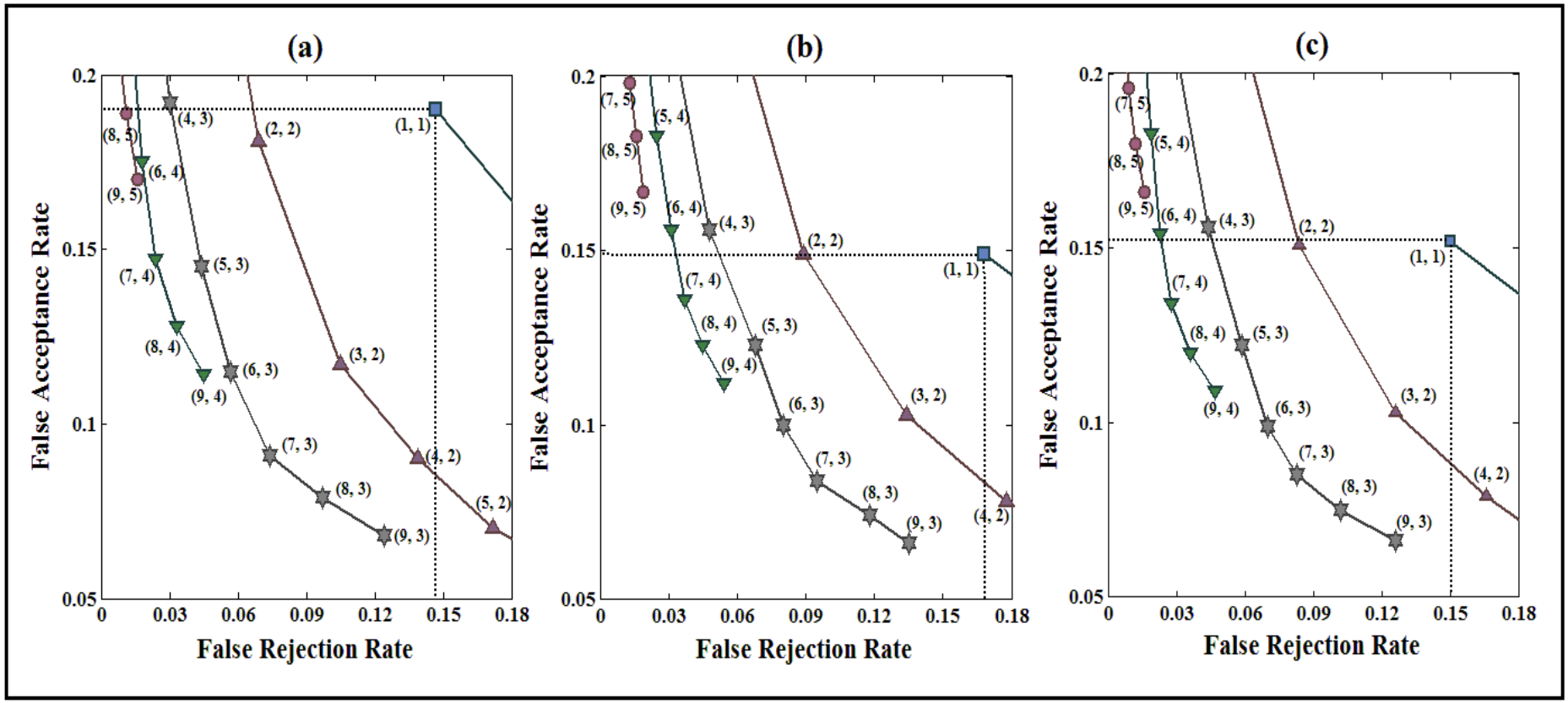

Figure 6.6 Verification Error Rates for multi-instance and multi-sample fusion of digits selected using (a) client dependent sequential error ratio, (b) impostor dependent sequential error ratio and (c) client-impostor dependent sequential error ratio (Each point on the curve represents FRR & FAR for fusion parameters (*n, m*)= (*number of instances, number of samples*). The points under the area *(1, 1)* represent the FRR and FAR values lower than base error *(1, 1)*)

dependent SER. Irrespective of the samples used for determining SER, the improvement in performance is achieved for parameter (*n*, *m*) combinations. The TER decreases with an increase in samples initially and then progressively increases when the increase in FAR is higher than decrease in FRR for the use of multiple samples (e.g., fusion of five samples results in TER higher than fusion of four samples).

For speaker verification based on random digit combinations, the accurate estimation of error rates is performed using the expressions for error rates (5.10) & (5.25) that incorporate correlation between the classifier decisions. The parameters required for this estimation are the base classifier errors and the variance in correlation coefficients obtained from known data. For speaker verification using predetermined digit combinations, lower error rates are obtained for fusion of correlated decisions than independent decisions when the dependence between decisions is favourable. Although, the classifier sequence selected using SER results in better fusion performance, the selected digit combinations may not ensure favourable dependence between the decisions - i.e., the ideal errors might be lower than the experimental errors. The ideal error rates are calculated using expressions for error rates developed with an independence assumption between classifier decisions. The ideal FRRs are higher than experimental FRRs (similar to fig. 4.26(I)) whereas ideal FARs is lower than experimental FARs (similar to fig. 4.26(II)). This difference in ideal and experimental error rates for the selected classifiers increases with digits and decreases with samples used for fusion. However, the fusion can be catastrophic with more number of samples because of a greater increase in false accepts than the decrease in false rejects.

Instead of increasing samples for reducing the difference between ideal and experimental error rates, the alternative is to select (using *SER*) a classifier at each stage that is favourably dependent with the decisions of previously selected classifier. The dependence on the client and impostor decisions is determined using expressions (5.35) and (5.36) respectively (chapter 5). This method can result in the fusion error rates slightly higher than selection using *SER*, if the actual sequence selected using *SER* is not a favourable digit combination. The speaker verification performance is improved by using user-dependent parameters such as speaker-specific thresholds, speaker-specific digit combinations and class-dependent measures such as favourable dependence and *Sequential Error Ratio* criteria. In addition, the combination of digits selected using SER can be utilized in the selection of a speaker-specific password or pin number (knowledge-based information) and this provides another factor (other than biometric information) in a multi-factor authentication scheme.

The proposed multi-instance and multi-sample fusion architecture is thus analytically and empirically shown to provide a control trade-off between false rejects and false accepts with simultaneous reduction in both error rates. The sequential architecture that integrates multiple instances with multiple samples is desirable in most of the speaker verification applications such as remote authentication, telephone and internet shopping applications. The tuning of parameters - the number of instances and samples - serve both security and user convenience requirements of speaker-specific verification. The architecture has the potential to improve the performance of even weaker classifiers when the decisions from multiple instances and/or samples are favourable. The architecture investigated here using text-dependent speaker verification is applicable to verification using other biometric modalities such as handwriting, fingerprints and key strokes.

6.6 Conclusion

The design of a multiple classifier system has been shown to have significant effect on the fusion performance. An efficient design requires the optimization of the combination method and then selection of optimal classifies. The proposed combination method, sequential fusion of multiple instances (*'AND Rule'*) and multiple samples (*'OR Rule'*), is pre-defined for the improvement in performance. As combining all available classifiers has been shown to be expensive and rarely optimal, the various classifier selection criteria are investigated for the proposed fusion scheme.

Given a fixed set of classifiers, the classifier selection based on *'Best Combination Performance*' rule provides the best possible performance for classifier combinations. However, this heuristic rule requires exhaustive evaluations and the search complexity increases with instances. The simple diversity and minimum weighted error measures have been shown to have weak correlation with combination performance. Although, the double fault, measure of difficulty and adaboost measures are good evaluation criteria, the differences in fusion performance compared to *'Best Combination Performance*' rule are high.

A new selection criterion - *sequential error ratio (SER)* - specifically tuned to the characteristics of sequential *'AND fusion'* was proposed for the classifier selection. The empirical evaluations of proposed architecture have shown that better performances for digit-combinations were achieved for the *SER* measure compared to other selection criteria. The number of classifier comparisons required at any stage of selection using *SER* was equal to

the number of classifiers remaining in the classifier set. As *'best combination performance'* measure requires exhaustive evaluations, the '*Sequential Error Ratio'* was shown as a better alternative for obtaining close to the best possible performance with reasonable complexity. Though the evaluations of the proposed measure in this paper are presented for text-dependent speaker verification, the sequential error ratio criterion is also applicable to the multibiometric architecture of other modalities such as fingerprint and handwriting samples.

The proposed fusion architecture is shown to reduce both false rejects and false accepts simultaneously. The performance of this sequential method is demonstrated, in the dissertation, to be dependent on the order of combination and the number of digits (instances) used for verification. In addition, the verification error rates are better predicted by incorporating *user-dependent* parameters such as speaker-specific thresholds, speaker-specific digit combinations and *class-dependent* measures such as favourable dependence and *Sequential Error Ratio* criteria.

Chapter 7
Conclusions and Future Directions

This chapter provides a summary of the work presented in this dissertation and conclusions drawn. The major problem addressed here concerns to the reliability of the performance of a biometric identity verification system. Fusion performance degrades because of the inter-user similarity (increases false accept) and the intra-user variability (increases false rejects). Therefore, error rates for the state-of-the-art biometric systems still cannot identify individuals with complete accuracy. Although accuracy improves with the combination of multiple sources of biometrics, both the verification error rates are difficult to reduce simultaneously.

7.1 Conclusions

The focus of this dissertation is to investigate a method that better controls the trade-off between verification error rates. Irrespective of the modality used for verification, the performance of fusion is improved with appropriate design architecture. Therefore, a sequential architecture is proposed in this dissertation for reducing both false rejects and false accepts with trade-off in verification time. Text-dependent speaker verification platform is used for analysing this proposed architecture that is applicable to verification from spoken digit strings, such as credit card numbers in telephone or voice over internet protocol based applications. The main contributions from this dissertation are in three areas of classifier combination. The next sections present the findings and future work in these three areas of multibiometrics.

7.1.1 Classifier Fusion Architecture

A multibiometric system design depends on various factors such as sources of information, acquisition and processing architecture, level of information fusion and methodology. As a decision is made in serial processing approach without waiting for all the available biometrics information, the sequential fusion method offers control over the trade-off between performance and amount of data required for making a reliable decision. The proposed multibiometric architecture considered the ***sequential integration*** of multiple

instances and multiple samples using *'AND and OR rules'* respectively. Analytical expressions of error rates were derived separately for multi-instance fusion and multi-sample fusion schemes where classifier decisions were assumed statistically independent. The error rates for the proposed fusion were then obtained by combining the *expressions for multi-instance and multi-sample fusion* schemes [32].

The experimental analysis of the sequential architecture was performed for text-dependent speaker verification using HMM based digit dependent speaker models. The sequential fusion of multiple instances reduced the number of false accepts at the cost of an increase in false rejects compared to base classifiers. The sequential fusion of multiple samples was shown to be complementary to multi-instance fusion. For multi-sample fusion, the number of false rejects was lower and the false accepts were higher compared to the base classifier (instance) performances. The decrease in FRR or the increase in FAR was higher for adaptive nature of samples rather than random repetitions [33]. The ***sequential fusion*** of decisions from multiple instances and samples was empirically shown to have a better control over the ***trade-off between false rejects and false accepts***. The fusion method has the potential to reduce both FRR and FAR simultaneously even for weak classifiers with relatively low performance [32]. Further, the sequential fusion scheme evaluated can be employed as an effective *anti-spoofing method* where the increase in false accepts because of impostor adaptive samples, generated using voice conversion techniques, was reduced through appropriate fusion of multiple instances and multiple samples [33].

In order to better explain the trade-off between error rates, the proposed architecture is applied for data obtained using random generation of binary values where '1' and '0' represents an acceptance and a rejection decision respectively. The proposed fusion scheme is evaluated for the combination of 30 classifiers each with 10,000 client and impostor decisions. The base classifiers are with mean EER of 3% with the individual error rates less than 5% for each classifier. As the outputs at the decision level are either '1' or '0', the results obtained can be representative of fusion method irrespective of the biometric modality or the feature extraction and classification algorithms used for verification.

When the base classifiers are of state-of-the-art performance with low error rates (suppose < 5%), the classifier combination results in error rates comparable to state-of-the-art fusion performances. For example, the FRR reduces from 3% to 0.02% whereas FAR reduces from 3% to 0.01% for the fusion of seven classifiers with three samples at each classifier ((7, 3) in fig. 7.1(a)). Whereas the fusion of three classifiers with three repeated samples (3, 3)

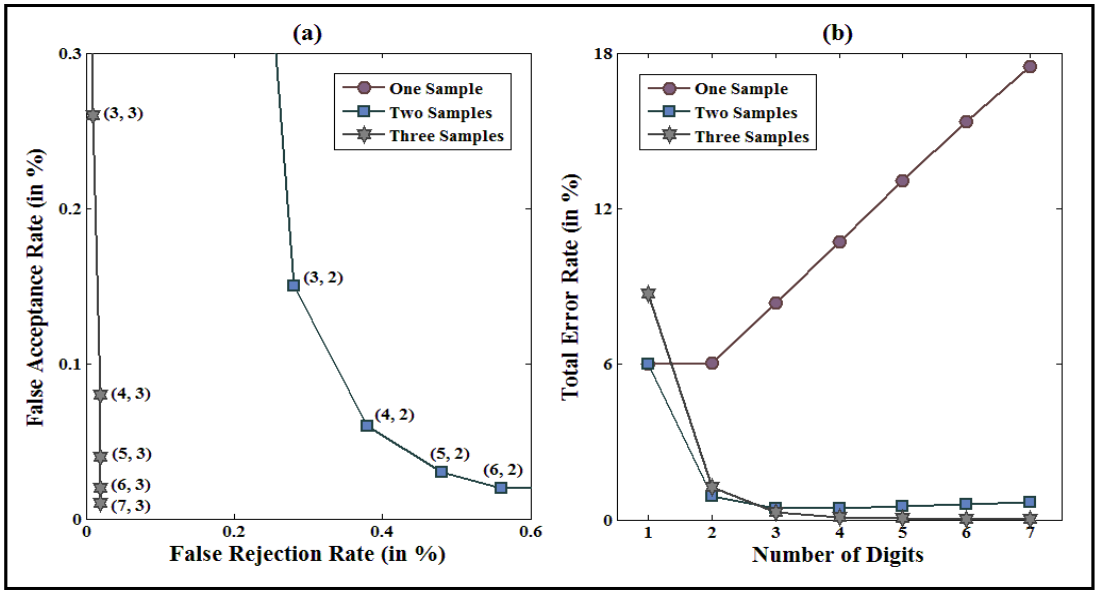

Figure 7.1 Error Rates for fusion of *'n'* instances and *'m'* samples (a) FRR vs. FAR (b) TER

reduces the false rejects and false accepts to 0.01% and 0.03% respectively. Therefore, the trade-off between false accepts and false rejects is better controlled by the ***tuning of the parameters***, '*n*' classifiers/instances and '*m*' attempts/samples. It is also demonstrated that the improvement in fusion performance is dependent on the ***base classifier performances*** and the trade-off is achieved irrespective of the modality and classification methods used for verification. The proposed sequential architecture was also demonstrated to achieve better performance (lower total error rates) compared to other fusion rules such as *MAX*, *MIN*, *MEDIAN*, *OR*, *AND*, and majority voting.

The theoretical or ideal error rates (calculated with an assumption of independence between the decisions) were supposed to be equal to experimental error rates for tests performed on text-dependent speaker verification. However, there existed a significant difference between ideal and experimental error rates (*FRR* and *FAR*) for fusion of multiple instances and samples [32]. One likely reason for this difference in error rates was the statistical dependence between the classifier decisions [36]. The input data (sample for an instance) presented at each decision stage for text-dependent speaker verification can thus be correlated even though the text is different.

7.1.2 Classifier Correlation Modelling

The assumption of independence holds good for combination of different biometric characteristics (e.g., face and voice) but may not be true for multibiometrics from a single

modality. Therefore, the expressions developed for fusion of independent decisions [32] are modified to incorporate the dependence between the classifier decisions [36]. The exact class-conditional error rates for the fusion of correlated decisions were estimated using the full expansion of *Bahadur-Lazarsfeld Expansion* (BLE). The error rates for multi-instance fusion were developed considering the conditions of acceptance from each of '*n*' decisions. Similarly, the error rates for multi-sample fusion were expressed using the BL expansion and the vector of rejections from multiple samples. The expressions from these fusion schemes were integrated for the proposed fusion, i.e., the multi-sample fusion error rates were substituted as base errors in the expressions for multi-instance fusion error rates [36]. The error rates for the unseen data (test dataset) were predicted using the base error rates and variance in correlation coefficients from known data (different tune datasets) that considers all the (prior) conditions under which a user may be tested. When there is an overlap in correlation for digit-combinations of tune and test datasets, maximum and minimum error rates (i.e., error bounds) were estimated for fixed '*n*' and '*m*' values using derived equations.

For statistically dependent classifier decisions, the error rates after decision fusion were larger or smaller than when the classifier decisions are *statistically independent*. The dependence is '*favourable*' when the error rates after fusion using either *'AND Rule'* or *'OR rule'* were smaller than those of independent classifier decisions. The *ideal false rejection rates* (FRR) were observed to be higher than the *experimental/predicted FRRs* whereas the ideal *false acceptance rates* (FAR) were lower than the *experimental/predicted FARs* for multi-instance fusion. The results for multi-sample fusion are complementary for multi-instance fusion [36]. The correlation coefficients for random samples were small and the mean error rates for fusion of independent and dependent decisions from random samples were similar, whereas, for adaptive samples the difference between these error rates was significant [33]. Thus, the type of repetitive presentation, i.e., random or adaptive sample, was differentiated using correlation between the decisions.

As the correlation coefficient for an '*n*'th instance is dependent on the previous on 2^{nd}, 3^{rd} ..., $(n-1)$th order coefficients, its relationship with error difference (between ideal & predicted error rates) is not direct. In the fusion literature, one sufficient condition for analysing the dependence between '*n*' classifiers was presented based on signs of the correlation coefficient. However, the expressions were developed, in the dissertation, for determining the conditions of favourable dependence between '*n*' correlated classifier decisions [40]. The dependence between the classifier decisions was determined using the

parameters - base error rates and magnitude of correlation between decisions. Improvement in performance was achieved irrespective of the number of samples and instances used for fusion when client-impostor favourable combinations were used for verification. With client-impostor favourable combinations, the experimental error rates (fusion of dependent decisions) were always lower than ideal error rates (fusion of independent decisions). These digit combinations were mostly different between speakers but were similar for a speaker across different datasets. Therefore, use of speaker-specific digit combinations was considered as another measure to ensure reliable identity verification.

In the most general case, $2\left(2^n - n - 1\right)$ correlation coefficients were required for prediction of error rates for *'n'* instances/samples. With the increase in instances/samples used for fusion, the computations required for calculation of correlation increases in exponential order. When only the most important correlation coefficients are incorporated, the empirical evaluation of the proposed fusion scheme had demonstrated that accurate estimation of errors was achieved using only ***second*** and ***third-order coefficients***. The theoretical analysis of the favourable/unfavourable dependence presented in this dissertation can be extended to investigate the necessary conditions for determining the optimal dependence (best/worst dependence) between '*n*' classifier decisions.

7.1.3 Optimal Classifier Selection

The design of a multibiometric system has a significant effect on the fusion performance. As combining all available classifiers in a large pool is expensive and rarely optimal, classifier selection methods were investigated for determining a classifier at each stage of fusion for best possible performance. As the efficiency of classifier selection method depends on the selection criterion, the fusion performance for selection based on performances and diversity measures were evaluated on text-dependent speaker verification. When '*k*' best classifiers were selected using *'Choose k Best'*, the fusion of only the best-chosen classifiers resulted in lower error rates (*'n = k'*). When the combination method performance is itself used for classifier selection, best possible performance was achieved irrespective of the number of classifiers and their individual performances. At each decision stage, the best combination performance is empirically selected from the entire set of possible combinations, i.e., $(2^n - 1)$ comparisons are required for fusion at '*n*' stages. For dynamic classifier selection, the classifiers with high diversity or minimum weighted error were selected at each stage of fusion for best possible performance. Although the double fault,

measure of difficulty and adaboost measures were good evaluation criteria, the difference in fusion performances compared to *'minimum combination errors'* (MCE) measure or *'best combination performance'* heuristic rule was high.

A new selection criterion - *sequential error ratio (SER)* - that was specifically tuned to the characteristics of sequential *'AND fusion'* was proposed for the classifier selection. Therefore, the classifier with decisions that mostly agree rather than disagree with previous classifiers' correct decisions results in better overall performance. The classifier selected has the minimum ratio of the number of input samples on which the classifier disagrees with previous correct decisions to the number of input samples on which the classifier agrees with the previous correct decisions. The empirical evaluations of proposed architecture have shown that better performances for digit-combinations were achieved for the *SER* measure compared to other selection criteria. The number of classifier comparisons required at any stage of selection using *SER* was equal to the number of classifiers remaining in the classifier set. As *'best combination performance'* measure requires exhaustive evaluations in exponential order that increases with digits used in fusion, the '*Sequential Error Ratio'* was considered the better alternative for obtaining close to the best possible performance with reasonable complexity.

The *sequential error ratio* criterion considers only the decisions that either agrees or disagrees with all the previous classifiers' correct decisions. The classifiers selected thus results in false accepts greater than best possible whereas the false rejection rates were close to the best possible for each digit-combination. Irrespective of this difference in false accepts, the *SER* measure was observed to be better than other selection criteria for fusion of *'n'* instances and *'m'* samples. The order of the selection also had significance on sequential fusion performance. As the selected digit-sequences can be different between speakers but similar for a speaker across different datasets, the use of speaker-specific digit combinations enabled reliable identity verification.

For theoretical estimation of error rates on unseen data, the classifier set with best combination performance for an individual were pre-determined on the known data (tune dataset) and later used in text-dependent speaker testing of unseen data (test dataset). The *SER* based classifier sequence was used for better prediction of fusion performance on test dataset given the error rates for base classifiers and the variance in correlation coefficients from the tune dataset (known data) were also known. As the base classifiers for both the tune and test datasets were assumed to be similar in performance, the parameters (*n, m*) used to

control the trade-off between false rejects and false accepts on tune dataset were also applicable to test dataset. Although the fusion of digits selected using *SER* ensured near to best possible performance, the error rates obtained may not be lower than the fusion of independent decisions. When the classifier that is favourably dependent with the decisions of previously selected classifier were selected using *SER* at each stage of fusion, the fusion of dependent decisions resulted in lower error rates than fusion of independent decisions. Therefore, user-dependent parameters such as speaker-specific thresholds, speaker-specific digit combinations and class-dependent measures such as favourable dependence and *Sequential Error Ratio* selection criteria were shown to improve the fusion performance of speaker verification.

7.2 Summary of Original Contributions

The original research contributions from this dissertation are in three areas of multibiometric fusion design. These contributions are summarized as follows

1) A multibiometric fusion architecture was proposed that incorporate sequential combination of decisions from multiple instances and multiple samples for biometric identity verification.
2) Expressions were developed for each type of verification error rates, false rejection rate and false acceptance rate, for the proposed fused architecture were derived for the case of *statistically independent classifier decisions*.
3) Empirical evaluation of the proposed sequential fusion architecture using text-dependent speaker verification using HMM based digit dependent models has been demonstrated to better control over the *trade-off* between *false alarms* and *false rejects.*
4) Analytical and empirical analysis presented on the nature of repetitive samples - *random* and *adaptive,* emulates the case in real applications where a speaker tries to adapt to claimed model. The sequential fusion scheme was employed as an effective anti-spoofing method where the increase in false accepts was reduced through appropriate combination of multiple instances and multiple samples.
5) The expressions developed for fusion error rates were modified using *Bahadur-Lazarsfeld expansion* to incorporate correlation between the classifier decisions from multiple instances and multiple samples.

6) The equations for verification error rates were experimentally evaluated for fusion of correlated decisions from multiple instances and multiple samples. The FRR (FAR) for fusion of independent decision were higher (lower) than that of dependent decisions. The difference between ideal error rates (independent decisions) and predicated error rates (dependent decisions) and the correlation between decisions was shown to decrease with an increase in digits and samples used for fusion.
7) Theoretical analysis for determining the conditions of favourable/unfavourable dependence between '*n*' correlated classifier decisions for *'AND and OR Rules'* was presented.
8) The statistical dependence between classifier decisions was empirically evaluated for determining *favourable classifier combinations* with performance better than that of fusion of statistically independent decisions.
9) A new selection criterion - *sequential error ratio* - was proposed for determining classifiers that result in optimal performance at each decision stage.
10) Empirical evaluation was presented for incorporation of *user-dependent* (such as speaker-dependent thresholds and speaker-specific digit combinations) and *class-dependent* (such as the client-impostor dependent favourable combinations and FRR-FAR based threshold estimation) information for the proposed fusion scheme.

In summary, the proposed architecture enables control over the trade-off between verification error rates (false accepts and false rejects) for a multibiometric system that operates on (single or multiple) thresholds tuned by adaptation algorithms used for different biometrics. The expressions derived for these error rates using base classifier performances and the variance in correlation between the classifier decisions enables to better predict the false acceptance and false rejection rates for the fusion of multiple instances and multiple samples. It was also demonstrated that the performance of the sequential method is dependent on the number of classifiers, the order in which classifiers (instance) are combined and the nature of verification attempts (samples). The fusion performance is improved by incorporating *user-dependent* and *class-dependent* information.

The proposed architecture is desirable in most of the speaker verification applications such as remote authentication, telephone and internet shopping applications. The tuning of parameters of the architecture, the number of attempts at each decision stage (samples) and the number of decision stages (instances), serve both security and user convenience

requirements of speaker-specific verification. The architecture also enables multifactor authentication by combining the knowledge-based speaker verification (e.g., using credit card numbers, SSN or speaker specific passwords) with biometrics, thereby improving the accuracy of verification.

The multibiometric architecture proposed combines information at decision level and thus the fusion framework is generalized for different sources of biometrics without consideration to the type of biometric data processing and classification methods. Although the biometrics - instances and samples from a speaker are used for evaluation, the architecture is applicable to other sources of biometrics or even modalities. For example, a user can be verified using the proposed architecture with fusion of multiple instances (fingerprints - such as left and right index fingers) and multiple algorithms (different feature/classification algorithms for each instance).

7.3 Limitations and Future Directions

The architecture and analysis presented in this dissertation for text-dependent speaker verification is applicable to any other biometric modalities (such as handwriting, fingerprints) and it is a clear future area of research. The level of security and in turn the number of classifiers used for fusion depends on the application requirements. For isolated digit based systems, the number is close to ten (or multiples of this if multiple languages are included). As individual digits were used for text-dependent speaker verification performance, the number of instances used for verification is limited. The evaluation could be better analysed by considering the architecture for large classifier pool or spoken text/utterances of greater length (such as two connected digit combinations, or strings) that have a higher number of possible inputs. Although the number of classifiers available for fusion can be increased to order of 100s or more, the user cannot be expected to realistically spend all that much time for verification. Therefore, the task here is to select a subset of classifiers that when combined results in the best possible performance.

There are 2^{2^n} possible rules for fusion of *'n'* decisions and adaptive algorithms have been used to determine the optimal fusion rule and sensor operating points for achieving varying levels of security [250]. Although monotonic rules are ignored, the number of fusion rules to be analysed with each addition of an instance increases the complexity (Out of the 256 possible fusion rules for the fusion of three instances, the 20 non-monotonic rules need to

be analysed. This number increases with each addition of an instance). The limitations of this approach [250] are addressed where the adaptive combination of multiple biometric modalities is performed at the match score level for desired level of security [251]. Costs and selected rules need to be changed for each set of threshold or operating points. With the fusion rules *'AND'* and *'OR'* (used in the dissertation), the decision thresholds or operating points for individual classifier systems are fixed. This enables the estimation of the probability distributions for proposed fusion scheme using expressions for errors based on decision fusion. These equations may not be as straightforward with score fusion as the final error will be a function of the threshold in the joint probability density space and would be a multi-dimensional integral in general. Further, the correlation between the statistically dependent decisions is used for determining the probability density function using the Bahadur-Lazarsfeld Expansion (BLE) approach. The expansion applies to the simplified case of binary vectors (1-accept and 0-reject) for decision fusion and therefore the expansion is simpler for decisions rather than scores. The analysis similar to BLE can be considered using score correlations of higher than second order that delves into the domain of higher order statistics with continuous variables.

The prediction of error rates for fusion of *'n'* instances/samples, in the most general case, requires $2\left(2^n - n - 1\right)$ correlation coefficients. With the increase in instances/samples used for fusion, the computations required for calculation of correlation increases in exponential order. Therefore, only the most important correlation coefficients are incorporated for accurate estimation of errors. Analytical methods can be derived in future to determine approximately the range over which any *'i'*th order (i =2, 3, 4, 5...n) coefficients is neglected, i.e. the appropriate truncation of Bahadur-Lazarsfeld Expansion to avoid unrealistic error rates.

The proposed architecture is directly applicable to security-based transactions using speaker verification from spoken digit strings such as credit card numbers in telephone or voice over internet protocol based applications. With this fusion scheme, both the verification error rates are reduced with a trade-off in the time required for a verification transaction. This increased time, in an application scenario such as an online transaction from a home or office computer, will be borne by the claimant and thus need not hold up other system users or be added on as additional cost to the service provider. This is especially so, if the processing is done locally and only results are communicated. The models required may need to be downloaded but that is a one-off operation.

Apart from the remote verification applications that require an individual to be verified rather than identified, the proposed architecture could be applicable for speaker identification applications involving monitoring, surveillance and automated ID tagging. With closed-set speaker identification, the unknown speaker's utterance is compared against all speaker models in the pool and the performance (accuracy and complexity) is degraded with an increase in the number of speakers in the pool. In open-set speaker identification, the speaker to be identified can be from the general population. As it is not possible to identify arbitrary people (speakers without models), if the speaker to be identified does not belong to the pool or database of known speakers then the speaker can be rejected, otherwise a random individual from the pool will be identified. With the proposed architecture, most of the speakers in the pool could be rejected at early stages without having to identify the speaker for all the available instances. Therefore, serial combination of identification decisions reduces the false positive identification rate (FPIR) and the availability of multiple samples reduces the false negative identification rate (FNIR). However, the effectiveness of the proposed multi-sample and multi-instance fusion scheme in achieving controlled trade-off between the identification error rates can be evaluated experimentally in the future.

The design of the proposed architecture can be modified to incorporate the application system requirements. For *cost-driven applications*, the increase in system cost is limited when biometrics are acquired and processed without changing the existing technology. However, multiple biometrics (instances/samples/features) of the same modality may have dependent information and not perform equally. Although, performance improves with an increase in multiple biometrics the complexity of the recognition system increases leading to higher cost, longer recognition time and user inconvenience. For *performance-driven applications* with less consideration for the cost of system, the biometric sources employed for fusion could be different physically related modalities. For example, multiple instances (e.g., word or digit utterances) from a speaker can be sequentially combined using *'AND fusion'*. When a speaker is rejected at an instance, the decision from the verification of lip reading or handwriting samples can be combined with that of voice using *'OR rule'*. For this architecture, the system may require individual sensors, different feature extraction and modelling algorithms for each modality. Nevertheless, the advantage of using multiple biometric traits is that independent information is made available at each stage of fusion.

Appendix A

A.1 VTLN-Based Voice Conversion

Voice conversion is used to convert speech of a source speaker into target speaker, using a transformation function, by replacing the physical characteristics of voice without altering the message contained in speech [178]. As explained in section 3.4.2.2, Vocal tract length normalization (VTLN) based voice conversion is used in this dissertation for obtaining additional speech data. VTLN [179] tries to compensate for the effect of speaker-dependent vocal tract lengths by warping the frequency axis of phase and magnitude spectrum [252]. The same technique can be used for the modification of a source speaker's voice in order to sound like another speaker.

VTLN can be applied to either frequency or time spectrum. The concatenation of speech segments and the prosodical manipulation (intonation, speaking rate, etc.) are often based on TD-PSOLA (time domain pitch-synchronous overlap and add). The application of FD-VTLN to speech synthesis requires the transformation from time to frequency domain and the other way around using DFT (discrete Fourier transformation) and inverse DFT, respectively (fig. A.1). However, the processing resources required for conversion can be limited using the TD-VTLN (time domain VTLN).

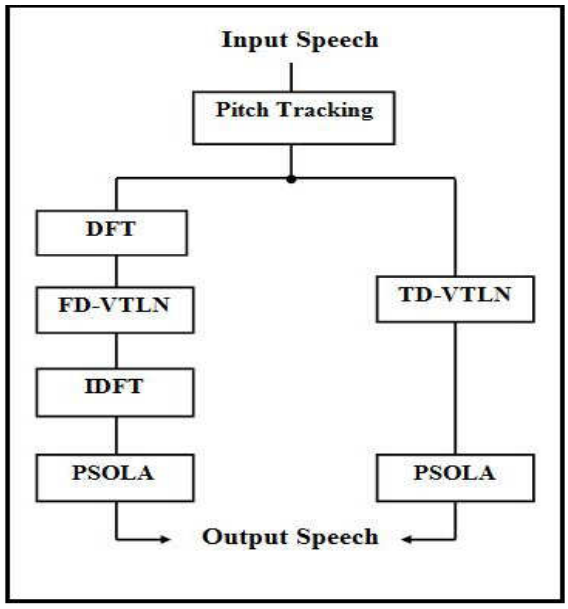

Figure A.1 Frequency Domain and Time Domain VTLN based Voice Conversion

A frame is composed of two pitch periods, the frame overlap is one period i.e., when $p_1^{M+1} = p_1, p_2, ...p_{M+1}$ is the sequence of considered pitch periods then (A.1) is the sequence of frames. This avoids signal discontinuities in the overlap-and-add concatenation.

$$x_1^M = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \begin{pmatrix} p_2 \\ p_3 \end{pmatrix}, ..., \begin{pmatrix} p_M \\ p_{M+1} \end{pmatrix} \tag{A.1}$$

- **Time domain PSOLA**: Given the time waveform of the frames to be processed, a Hamming window is applied,

$$x^1(t) = \frac{x(t)}{2}\left(1 - \cos\left(2\pi \frac{t}{T}\right)\right), 0 \le t \le T \tag{A.2}$$

moreover, the frames are overlapped. Here, T is the frame's duration. Each frame consists of two pitch periods, and the standard overlap is one period. The overlapping period at the successive frames $x_m = \begin{pmatrix} p_m \\ p_{m+1} \end{pmatrix}$ and $x_{m+1} = \begin{pmatrix} p_{m+1} \\ p_{m+2} \end{pmatrix}$ is p_{m+1}. It is windowed by the falling half of the Hamming window for x_m and by the rising half for x_{m+1}. Adding together both contributions produces

$$P_{m+1}(t) = \frac{p_{m+1}(t)}{2}\left(1 - \cos\left(2\pi \frac{t + \frac{T}{2}}{T}\right)\right) + \frac{p_{m+1}(t)}{2}\left(1 - \cos\left(2\pi \frac{t}{T}\right)\right)$$

$$= \frac{p_{m+1}(t)}{2}\left(2 - \cos\left(2\pi \frac{t}{T} + \pi\right) - \cos\left(2\pi \frac{t}{T}\right)\right) \qquad = p_{m+1}(t) \tag{A.3}$$

This leaves the signal unchanged in the standard case and, consequently, does not produce additional distortions. However, this does not apply if the fundamental frequency or the timing is changed. This is done by shifting the overlapping frames so that the resulting period length becomes

$$\overline{T} = \frac{T}{\rho} \;\; with \;\; \rho = \frac{\overline{f_{0,t}}}{f_{0,s}} \tag{A.4}$$

Where $\overline{f_0}$ is the mean fundamental frequency observed in the training data and s and t denote source and target, respectively. In order to keep or consciously change the temporal evolution of the speech signal, time frames have to be repeated or deleted, respectively.

- **Frequency domain PSOLA**: If the considered frames are given in frequency domain, i.e. as magnitude and phase spectra, interpolation can be applied on to them to match the target number of spectral lines given by equation A.3.3. If time domain is considered using inverse discrete Fourier transformation, the standard case introduced for time domain PSOLA applies and additional signal distortion is negligible. Here, the main signal deterioration is due to the interpolation.

The pitch-synchronous frames can be extracted from a given speech signal. In voiced regions, the frame lengths depend on the fundamental frequency, in unvoiced regions, the pitch extraction algorithm utilizes a mean approximation. By applying DFT without zero padding to the frames, complex-valued spectra with distinct numbers of spectral lines can be obtained. The realization that the warping of frequency axis of the magnitude spectrum can lead to a considerable performance gain resulted in several warping functions. They can be distinguished regarding the number of parameters describing the particular function and their linearity or nonlinearity, respectively. In general, a warping function is defined as

$$\overline{w}\left(w \mid \varepsilon_1, \varepsilon_2 ....\right);\ 0 \le w,\ \overline{w} \le \pi \tag{A.5}$$

Where $\varepsilon_1, \varepsilon_2, ...$ are the warping parameters and w is the normalized frequency with π corresponding to half the sampling frequency according to the Nyquist criterion. When VTLN is applied to voice conversion, different warping function results in very similar spectra. Here the piece-wise linear warping function with several segments that includes the two-segment function is considered [252]:

$$\overline{w}\left(w \mid w_1^t, w_1^{-I} = \alpha_i w + \beta_i\right)\ for\ w_i \le w \le w_{i+1} \tag{A.6}$$

$$with\ \ \alpha_i = \frac{\overline{w}_{i+1} - \overline{w}_i}{w_{i+1} - w_i} \tag{A.7}$$

$$\beta_i = \overline{w}_{i+1} - \alpha_i w_{i+1} \tag{A.8}$$

$$0 = w_0 < w_1 < ..... < w_I < w_{I+1} = \pi \quad for \quad \overline{w}_i \quad equivalent$$

Both frequency domain and time domain VTLN based techniques are used for conversion in this work. The system properties used for training and estimation of parameters using the FD-VTLN and TD-VTLN based voice conversion are given in chapter 3 (table 3.4).

A.2 Scott-Knott procedure

1. The classifiers are sorted in ascending order based on mean error rates

2. The sorted classifiers are separated into two groups $E_1 = \{\bar{e}_1 .. \bar{e}_i\}$ and $E_1 = \{\bar{e}_{i+1} .. \bar{e}_n\}$

3. The sum of squares for classifiers between-groups is computed

$$B_i = k(|E_1|\left(\bar{e}_{E_1} - \bar{e}_E\right)^2 + (|E_2|(\bar{e}_{E_2} - \bar{e}_E)^2$$

Where $|E_1|$= *i* and $|E_2|$ = (*n* - *i*) and $\bar{e}_E, \bar{e}_{E_1}, \bar{e}_{E_2}$ are the means of groups E, E_1 andE_2:

$$\bar{e}_E = \frac{1}{n}\sum_{i=1}^{n}\bar{e}_i \qquad \bar{e}_{E_1} = \frac{1}{|E_1|}\sum_{i\in E_1}\bar{e}_i \qquad \bar{e}_{E_2} = \frac{1}{|E_2|}\sum_{i\in E_2}\bar{e}_i$$

4. The parameters obtained in step 3 are used to determine the classifier partition that maximizes the sum of squares

$$B_{i*} = max\left\{k(|E_1|\left(\bar{e}_{E_1} - \bar{e}_E\right)^2 + |E_2|\left(\bar{e}_{E_2} - \bar{e}_E\right)^2)\right\}$$

5. The statistics $s^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(e_{ij}-\bar{e}_E)^2}{nk}$ and $\lambda = \frac{\pi}{2(\pi-2)}\frac{B_{i*}}{s^2}$ are computed which has approximately a X_v^2 distribution where the degrees of freedom are given by v = n/(π-2) (rounded).

6. If $\lambda > X_{v:\alpha}^2$ (where α is a predefined level), then set n =$|E_1|$, E = E_1 and return to step 1 to repeat the steps with the first group with the smallest means. If $\lambda < X_{v:\alpha}^2$ then all the means fall in the same homogeneous group.

References

[1] A. Jain, A. Ross, and K. Nandakumar, *Handbook of multibiometrics*: Springer-Verlag New York Inc, 2006.

[2] K. Nandakumar, "Multibiometric systems: Fusion strategies and template security," PhD thesis, Department of Computer Science and Engineering, Michigan State University, 2008.

[3] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of biometrics*: Springer, 2008.

[4] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft Biometric Traits for Personal Recognition Systems," in *Biometric Authentication*. vol. 3072, D. Zhang and A. K. Jain, Eds., ed: Springer Berlin / Heidelberg, 2004, pp. 1-40.

[5] D. Mahoni, D. Maio, and A. Jain, "Handbook of fingerprint recognition," *New York: Springer-Vedag*, p. 173—202, 2003.

[6] J. Ashbourn, *Biometrics: advanced identity verification*. London: Springer-Verlag, 2000.

[7] J. Benesty, M. Mohan Sondhi, and Y. Huang, "Text-Dependent Speaker Recognition," in *Springer handbook of speech processing*, ed: Springer, 2008.

[8] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 1895–1898.

[9] A. K. Jain, "Biometric authentication," *Scholarpedia*, vol. 3, p. 3716, 2008.

[10] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction to evaluating biometric systems," *IEEE Computer*, vol. 33, pp. 56-63, 2000.

[11] A. K. Jain and S. Prabhakar, "Can multi-biometrics improve performance?," in *Proceedings of the IEEE Workshop on Automatic Identification Advanced Technologies (AutoID)*, New Jersey, 1999, pp. 59-64.

[12] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 955-966, 1995.

[13] Z. Akhtar and N. Alfarid, "Robustness of Serial and Parallel Biometric Fusion against Spoof Attacks," in *Computer Networks and Intelligent Computing*. vol. 157, K. R. Venugopal and L. M. Patnaik, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 217-225.

[14] A. Jain, L. Hong, and Y. Kulkarni, "A multimodal biometric system using fingerprint, face and speech," in *Second International Conference on AVBPA,*, Washington D.C., USA, 1999, pp. 182-187.

[15] A. Ross, A. Jain, and J. Z. Qian, "Information fusion in biometrics," in *Proc. of 3rd International Conference on Audio- and Video-Based Person Authentication (AVBPA)*, Sweden, 2001, pp. 354-359.

[16] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226-239, 1998.

[17] K. Nandakumar, C. Yi, S. C. Dass, and A. K. Jain, "Likelihood Ratio-Based Biometric Score Fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, pp. 342-347, 2008.

[18] K. Takahashi, M. Mimura, Y. Isobe, and Y. Seto, "A secure and user-friendly multimodal biometric system," in *Biometric Technology for Human Identification.* vol. 5404, K. J. Anil and K. R. Nalini, Eds., ed: Proceedings of the SPIE, 2004, pp. 12-19.

[19] E. Vildjiounaite, S. M. Makela, M. Lindholm, V. Kyllonen, and H. Ailisto, "Increasing security of mobile devices by decreasing user effort in verification," in *Second International Conference on Systems and Networks Communications*, 2007, p. 80.

[20] A. Martin and M. Przybocki, "The NIST 1999 speaker recognition evaluation-an overview," *Digital Signal Processing,* vol. 10(1-3), pp. 1-18, 2000.

[21] A. C. Surendran, "Sequential Decisions for Faster and More Flexible Verification," in *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH),* Aalborg, Denmark, 2001, pp. 763-766.

[22] R. Kashi and W. Nelson, "Signature verification: benefits of multiple tries," in *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition*, 2002, pp. 424-427.

[23] V. Chandran and A. Nguyen, "Biometrics: New Perspectives on Multimodal and Client-Centered Systems," in *Proceedings of International Workshop on Recent Advances in Biometric Systems,* Kanpur, India, 2005, pp. 77-90.

[24] T. K. Ho, "Multiple classifier combination: Lessons and next steps," *Hybrid methods in pattern recognition,* vol. 74, pp. 171-198, 2002.

[25] F. Roli and G. Giacinto, "Design of multiple classifier systems," *Series in Machine Perception and Artificial Intelligence,* vol. 47, pp. 199-226, 2002.

[26] M. C. Cheung, K. K. Yiu, M. W. Mak, and S. Y. Kung, "Multi-sample fusion with constrained feature transformation for robust speaker verification," in *Proc. of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP04),* Montreal, Canada, 2004, pp. 1813-1816.

[27] S. Bengio, "Multimodal speech processing using asynchronous Hidden Markov Models," *Information Fusion,* vol. 5, pp. 81-89, 2004.

[28] Ramli D A, Rani N C, and I. K. A, "A Multi-Instance Speech Signal Data Fusion Approach for Biometrie Speaker Authentication System Enhancement," *World Applied Sciences Journal,* vol. 10, pp. 847-852, 2010.

[29] C. Sanderson and K. K. Paliwal, "Joint cohort normalization in a multi-feature speaker verification system," in *10th IEEE International Conference on Fuzzy Systems*, 2001, pp. 232-235.

[30] S. Sharma, H. Hermansky, and P. Vermuulen, "Combining information from multiple classifiers for speaker verification," in *Proc. Speaker Recogniton and Its Commercial and Forestic Applications Workshop (RLA2C),* Avignon, France, 1998, pp. 115-119.

[31] A. Ross and N. Poh, "Multibiometric Systems: Overview, Case Studies, and Open Issues," in *Handbook of Remote Biometrics*, M. Tistarelli, *et al.*, Eds., ed: Springer, 2009, pp. 273-292.

[32] V. P. Nallagatla and V. Chandran, "Sequential decision fusion for controlled detection errors," in *13th International Conference on Information Fusion*, Edinburgh, 2010.

[33] V. Nallagatla and V. Chandran, "Sequential Fusion of Decisions from Adaptive And Random Samples for Controlled Verification Errors," in *11th International Conference on Information Science, Signal Processing and their Applications*, Canada, 2012, pp. 793-798.

[34] N. Karthik, A. Ross, and A. K. Jain, "Biometric Fusion: Does Modeling Correlation Really Matter?," in *Proceedings of IEEE Third International Conference on BTAS*, Washington DC, 2009, pp. 271-276.

[35] L. I. Kuncheva and L. C. Jain, "Designing classifier fusion systems by genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, pp. 327-336, 2000.

[36] V. Nallagatla and V. Chandran, "Sequential Fusion Using Correlated Decisions for Controlled Verification Errors," in *Computer Analysis of Images and Patterns*. vol. 6855, P. Real, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2011, pp. 49-56.

[37] K. Venkataramani and B. V. K. Vijaya Kumar, "Conditionally-dependent classifier fusion using AND rule for improved biometric verification," in *International Conference on Advances in Pattern Recognition*, , August 2005, pp. 277-286.

[38] K. Venkataramani and B. V. K. V. Kumar, "OR rule fusion of conditionally dependent correlation filter based classifiers for improved biometric verification," in *Proceedings of SPIE*, 2006, p. 62450A.

[39] K. Venkataramani, "Optimal classifier ensembles for improved Biometric Verification," Ph.D, Carnegiw Mellon University, 2007.

[40] V. Nallagatla and V. Chandran, "Sequential Fusion of Decisions with Favourable Dependence for Controlled Verification Errors," in *11th International Conference on Information Science, Signal Processing and their Applications*, Canada, 2012, pp. 259-264.

[41] D. Partridge and W. B. Yates, "Engineering multiversion neural-net systems," *Neural computation*, vol. 8, pp. 869-893, 1996.

[42] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Machine Learning*, vol. 51, pp. 181-207, 2003.

[43] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Information Fusion*, vol. 6, pp. 63-81, 2005.

[44] V. Nallagatla and V. Chandran, "Classifier Selection using Sequential Error Ratio Criteria for Multi-Instance and Multi-Sample Fusion," in *Proceedings of the 6th International Conference on Signal Processing and Communication Systems*, Gold Coast, Australia, 2012, pp. 496-503.

[45] L. Allano, B. Dorizzi, and S. Garcia-Salicetti, "Tuning cost and performance in multi-biometric systems: A novel and consistent view of fusion strategies based on the Sequential Probability Ratio Test (SPRT)," *Pattern Recognition Letters*, vol. 31, pp. 884-890, 2010.

[46] J. Lee, B. Moghaddam, H. Pfister, and R. Machiraju, "Finding optimal views for 3D face shape modeling," *International Conference on Automatic Face and Gesture Recognition*, pp. 31-36, 2004.

[47] D. A. Ramli, N. H. C. Rani, and K. A. Ishak, "Performances of Weighted Sum-Rule Fusion Scheme in Multi-Instance and Multi-Modal Biometrie Systems," *World Applied Sciences Journal*, vol. 12, pp. 2160-2167, 2011.

[48] S. Prabhakar and A. K. Jain, "Decision-level fusion in fingerprint verification," *Pattern Recognition*, vol. 35, pp. 861-874, 2002.

[49] J. Jang, K. Park, J. Son, and Y. Lee, "Multi-unit Iris Recognition System by Image Check Algorithm," in *Biometric Authentication*. vol. 3072, D. Zhang and A. Jain, Eds., ed: Springer Berlin / Heidelberg, 2004, pp. 1-16.

[50] H. Hill, P. G. Schyns, and S. Akamatsu, "Information and viewpoint dependence in face recognition," *Cognition*, vol. 62, pp. 201-222, 1997.

[51] U. Uludag, A. Ross, and A. A. Jain, "Biometric template selection and update: a case study in fingerprints," *Pattern Recognition*, vol. 37, pp. 1533-1542, 2004.

[52] A. El-Sallam and A. Mian, "Speech-Video Synchronization Using Lips Movements and Speech Envelope Correlation," *Image Analysis and Recognition*, pp. 397-407, 2009.

[53] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multimodal 2D+3D face biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 619-624, 2005.

[54] H. Kekre, V. Bharadi, V. Singh, V. Kaul, and B. Nemade, "Hybrid Multimodal Biometric Recognition Using Kekre s Wavelets, 1D Transforms & Kekre s Vector Quantization Algorithms Based Feature Extraction of Face & Iris," in *Proceedings on International Conference and workshop on Emerging Trends in Technology*, 2011, pp. 29-34.

[55] W. Fenghua and H. Jiuqiang, "Information fusion in personal biometric authentication based on the iris pattern," *Measurement Science and Technology*, vol. 20, p. 045501, 2009.

[56] A. Jain, K. Nandakumar, X. Lu, and U. Park, "Integrating faces, fingerprints, and soft biometric traits for user recognition," *Biometric Authentication*, pp. 259-269, 2004.

[57] R. W. Frischholz and U. Dieckmann, "BiolD: a multimodal biometric identification system," *Computer*, vol. 33, pp. 64-68, 2000.

[58] G. L. Marcialis, F. Roli, and L. Didaci, "Personal identity verification by serial fusion of fingerprint and face matchers," *Pattern Recognition*, vol. 42, pp. 2807-2817, 2009.

[59] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, pp. 2270-2285, 2005.

[60] S. Iyengar, L. Prasad, and H. Min, "Advances in Distributed Sensor Technology," ed: Englewood Cli. s, NJ, 1995.

[61] X. Liu and T. Chen, "Geometry-assisted statistical modeling for face mosaicing," in *International Conference on Image Processing*, 2003, pp. II-883-6 vol.3.

[62] A. K. Jain and B. Chandrasekaran, "39 Dimensionality and sample size considerations in pattern recognition practice," in *Handbook of Statistics*. vol. Volume 2, P. R. Krishnaiah and L. N. Kanal, Eds., ed: Elsevier, 1982, pp. 835-855.

[63] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters,* vol. 18, pp. 853-858, 1997.

[64] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE,* vol. 85, pp. 6-23, 1997.

[65] J. Byeungwoo and D. A. Landgrebe, "Decision fusion approach for multitemporal classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 37, pp. 1227-1233, 1999.

[66] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Decision Fusion for the Classification of Urban Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 44, pp. 2828-2838, 2006.

[67] L. Gee and M. Abidi, "Multisensor fusion for decision-based control cues," in *Proc. of SPIE on Signal Processing, Sensor Fusion, and Target Recognition IX*, 2000, pp. 249-257.

[68] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Biometric Score Fusion: Likelihood Ratio, Matcher Correlation and Image Quality," Technical Report MSUCSE-07-182007.

[69] Q. Tao and R. Veldhuis, "Optimal decision fusion for a face verification system," *Advances in Biometrics,* pp. 958-967, 2007.

[70] M. E. Liggins, D. L. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*: CRC, 2008.

[71] D. Zhang, F. Song, Y. Xu, and Z. Liang, "Decision Level Fusion," in *Advanced Pattern Recognition Technologies with Applications to Biometrics*, ed: IGI Global, 2009, pp. 328-348.

[72] F. Roli, Š. Raudys, and G. Marcialis, "An experimental comparison of fixed and trained fusion rules for crisp classifier outputs," *Multiple Classifier Systems,* pp. 203-206, 2002.

[73] K. Tumer and J. Ghosh, "Linear and order statistics combiners for pattern classification," *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, pp. 127-162, 1999.

[74] F. Roli and G. Fumera, "Analysis of linear and order statistics combiners for fusion of imbalanced classifiers," *Multiple Classifier Systems,* pp. 252-261, 2002.

[75] J. Kittler, M. Ballette, J. Czyz, F. Roli, L. Vandendorpe, F. Roli, and J. Kittler, "Decision Level Fusion of Intramodal Personal Identity Verification Experts Multiple Classifier Systems." vol. 2364, ed: Springer Berlin / Heidelberg, 2002, pp. 1-4.

[76] J. Daugman, "Combining multiple biometrics," *The Computer Laboratory, Cambridge University,* 2004.

[77] J. R. Movellan and P. Mineiro, "Robust Sensor Fusion: Analysis and Application to Audio Visual Speech Recognition," *Machine Learning,* vol. 32, pp. 85-100, 1998.

[78] N. Fox, B. O'Mullane, and R. Reilly, "Audio-Visual Speaker Identification via Adaptive Fusion Using Reliability Estimates of Both Modalities," in *Audio- and*

Video-Based Biometric Person Authentication. vol. 3546, T. Kanade, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 163-174.

[79] N. Fox and R. Reilly, "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features," in *Audio- and Video-Based Biometric Person Authentication*. vol. 2688, J. Kittler and M. Nixon, Eds., ed: Springer Berlin / Heidelberg, 2003, pp. 1059-1059.

[80] A. K. Jain, F. D. Griess, and S. D. Connell, "On-line signature verification," *Pattern Recognition,* vol. 35, pp. 2963-2972, 2002.

[81] T. Matsui, T. Nishitani, and S. Firui, "Robust methods of updating model and a priori threshold in speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 97-100 vol. 1.

[82] A. K. Jain and A. Ross, "Learning user-specific parameters in a multibiometric system," in *Proceedings of International Conference on Image Processing*, 2002, pp. I-57-I-60 vol.1.

[83] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 29, pp. 254-272, 1981.

[84] K. A. Toh, J. Xudong, and Y. Wei-Yun, "Exploiting global and local decisions for multimodal biometrics verification," *Signal Processing, IEEE Transactions on,* vol. 52, pp. 3059-3072, 2004.

[85] N. Poh, "Multi-system biometric authentication: Optimal fusion and user-specific information," Swiss Federal Institute of Technology in Lausanne (EPFL), 2006.

[86] M. Golfarelli, D. Maio, and D. Malton, "On the error-reject trade-off in biometric verification systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, pp. 786-796, 1997.

[87] K. Tumer and J. Ghosh, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition,* vol. 29, pp. 341-348, 1996.

[88] G. Fumera and F. Roli, "Analysis of error-reject trade-off in linearly combined multiple classifiers," *Pattern Recognition,* vol. 37, pp. 1245-1265, 2004.

[89] J. J. Clark and A. L. Yuille, *Data fusion for sensory information processing systems* vol. 105: Springer, 1990.

[90] N. Poh, S. Bengio, and J. Korczak, "A multi-sample multi-source model for biometric authentication," in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 375-384.

[91] K. W. Bowyer, K. I. Chang, P. J. Flynn, and X. Chen, "Face recognition using 2-d, 3-d, and infrared: Is multimodal better than multisample?," *Proceedings of the IEEE,* vol. 94, pp. 2000-2012, 2006.

[92] A. Wald, *Sequential analysis*. New York: John Wiley and Sons, 1947.

[93] R. Vogt and S. Sridharan, "Minimising Speaker Verification Utterance Length through Confidence Based Early Verification Decisions," *Lecture Notes in Computer Science,* pp. 5558/2009:454-463, 2009.

[94] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics,* vol. 16, pp. 117-186, 1945.

[95] M. E. Schuckers, E. H. Sheldon, and H. A. Hartson, "When enough is enough: early stopping of biometrics error rate testing," in *IEEE Workshop on Automatic Identification Advanced Technologies*, 2007, pp. 7-12.

[96] M. Wojnarski, "Absolute contrasts in face detection with adaBoost cascade," *Rough Sets and Knowledge Technology*, pp. 174-180, 2007.

[97] F. Wang, X. Yao, and J. Han, "Improving iris recognition performance via multi-instance fusion at the score level," *Chinese Optics Letters*, vol. 6, pp. 824-826, 2008.

[98] X. Wu, K. Wang, D. Zhang, and N. Qi, "Combining Left and Right Irises for Personal Authentication," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*. vol. 4679, A. Yuille, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 145-152.

[99] A. K. Jain, S. Prabhakar, and A. Ross, "Fingerprint matching: Data acquisition and performance evaluation," *Department of Computer Science, Michigan State University, East Lansing, Tech. Rep. MSU-CPS-99–14*, 1999.

[100] J. L. Dugelay, J. C. Junqua, C. Kotropoulos, R. Kuhn, F. Perronnin, and I. Pitas, "Recent advances in biometric person authentication," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ,* 2002, pp. IV-IV.

[101] M. Beattie, B. V. K. Kumar, S. Lucey, and O. Tonguz, "Combining Verification Decisions in a Multi-vendor Environment," in *Audio- and Video-Based Biometric Person Authentication*. vol. 3546, T. Kanade, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 277-290.

[102] M. C. Cheung, M. W. Mak, and S. Y. Kung, "Adaptive decision fusion for multi-sample speaker verification over GSM networks," in *Eurospeech*, 2003, pp. 1681-1684.

[103] C. Li, G. Su, Y. Shang, Y. Li, and Y. Xiang, "Face Recognition Based on Pose-Variant Image Synthesis and Multi-level Multi-feature Fusion Analysis and Modeling of Faces and Gestures." vol. 4778, S. Zhou, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 261-275.

[104] A. J. O'toole and H. Bulthoff, "Face recognition across large viewpoint changes," in *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR)*, Zurich, Switzerland, 1995, pp. 326-331.

[105] G. R. Doddington, "Speaker recognition - Identifying people by their voices," *Proceedings of the IEEE*, vol. 73, pp. 1651-1664, 1985.

[106] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, "A Survey on Automatic Speaker Recognition Systems," in *Signal Processing and Multimedia*. vol. 123, T.-h. Kim, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 134-145.

[107] J. Naik and G. Doddington, "Evaluation of a high performance speaker verification system for access control," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1987, pp. 2392-2395.

[108] Z. Saquib, N. Salam, R. Nair, and N. Pandey, "Voiceprint Recognition Systems for Remote Authentication-A Survey," *International Journal of Hybrid Information Technology*, vol. 4, 2011.

[109] J. M. Naik, L. P. Netsch, and G. R. Doddington, "Speaker verification over long distance telephone lines," in *International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 524-527 vol.1.

[110] C. Schmandt and B. Arons, "A Conversational Telephone Messaging System," *IEEE Transactions on Consumer Electronics*, vol. CE-30, pp. xxi-xxiv, 1984.

[111] A. Alexander, "Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions," *International Journal of Speech Language and the Law*, vol. 14, pp. 145-147, 2007.

[112] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, 2001, pp. 231-218.

[113] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. II-53-6 vol.2.

[114] C. Barras, S. Meignier, and J. L. Gauvain, "Unsupervised online adaptation for speaker verification over the telephone," *Odyssey, Toledo, Spain*, 2004.

[115] J. P. Campbell, Jr., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437-1462, 1997.

[116] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, pp. 89-106, 1991.

[117] H. Melin, "Automatic speaker verification on site and by telephone: methods, applications and assessment," PhD Thesis, KTH, Stockholm, , December, 2006.

[118] M. F. BenZeghiba and H. Bourlard, "User-customized password speaker verification based on HMM/ANN and GMM models," in *Proc. 2002 International Conference on Spoken Language Processing (ICSLP)*, Denver CO, USA, 2002, pp. 1325-1328.

[119] S. K. Gupta and M. Savic, "Text-independent speaker verification based on broad phonetic segmentation of speech," *Digital Signal Processing*, vol. 2, pp. 69-79, 1992.

[120] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993, pp. 391-394 vol.2.

[121] D. Sturim, W. Campbell, and D. Reynolds, "Classification Methods for Speaker Recognition," in *Speaker Classification I*. vol. 4343, C. Müller, Ed., ed: Springer Berlin / Heidelberg, 2007, pp. 278-297.

[122] L. Lam, "Classifier combinations: implementations and theoretical issues," *Multiple Classifier Systems*, pp. 77-86, 2000.

[123] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," Entropic Cambridge Research Laboratory, Cambridge, England, 1997.

[124] A. Rosenberg, J. Delong, C. Lee, B. Juang, and F. Soong, "The use of cohort normalized scores for speaker recognition," in *International Conference on Spoken Language Processing* Alberta, Canada, 1992, pp. 599-602.

[125] L. P. Heck and M. Weintraub, "Handset-dependent background models for robust text-independent speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 1071-1074 vol.2.

[126] P. Lockwood, C. Baillargeat, J. Gillot, J. Boudy, and G. Faucon, "Noise reduction for speech enhancement in cars: Non-linear spectral subtraction/kalman filtering," in *in Proc. Eurospeech*, 1991, pp. 83-86.

[127] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 357-366, 1980.

[128] J. Allen, "Cochlear modeling," *ASSP Magazine, IEEE*, vol. 2, pp. 3-29, 1985.

[129] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition* vol. 103: Prentice hall Englewood Cliffs, New Jersey, 1993.

[130] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 52-59, 1986.

[131] T. Kato and T. Shimizu, "Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. II-57-60 vol.2.

[132] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.

[133] S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the ARPA Human Language Technology Workshop*, Princeton NJ, 1994, pp. 307-312.

[134] L. P. Heck and N. Mirghafori, "On-line unsupervised adaptation in speaker verification," in *Sixth International Conference on Spoken Language Processing* 2000, pp. 454-457.

[135] X. Huang and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 150-157, 1993.

[136] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 1995, pp. 676-679 vol.1.

[137] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer speech and language*, vol. 9, p. 171, 1995.

[138] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 39, pp. 806-814, 1991.

[139] P. Nguyen, "Fast speaker adaptation," *Industrial Thesis Report, Institut Eurécom*, 1998.

[140] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," in *International Conference on Acoustics, Speech, and Signal Processing*, 1995, pp. 680-683 vol.1.

[141] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[142] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 1997, pp. 963-966.

[143] J. Lindberg, J. Koolwaaij, H. Hutter, D. Genoud, M. Blomberg, J. Pierrot, and F. Bimbot, "Techniques for a priori decision threshold estimation in speaker verification," in *Proceedings of Speaker Verification RLA2C*, Avignon, 1998, pp. 89–92.

[144] J. Saeta and J. Hernando, "Weighting Scores to Improve Speaker-Dependent Threshold Estimation in Text-Dependent Speaker Verification," *Nonlinear Analyses and Algorithms for Speech Processing*, pp. 81-91, 2005.

[145] L. Heck, "On the deployment of speaker recognition for commercial applications: Issues and best practices," in *Proc. Odyssey Speaker Recognition Workshop*, 2004.

[146] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high-and low-level features for speaker recognition," in *Proc. of EuroSpeech*, 2003, pp. 2665–2668.

[147] C. Sanderson and K. K. Paliwal, "Information fusion for robust speaker verification," in *Proc. Eurospeech'01*, Scan-dinavia, 2001.

[148] K. R. Farrell, "Text-dependent speaker verification using data fusion," in *Proceedings ICASSP*, Detroit, MI, 1995, pp. 349-352

[149] M. Przybocki and A. Martin, "NIST speaker recognition evaluation chronicles," in *Odyssey Workshop*, 2004, pp. 12–22.

[150] D. Boies, M. Hébert, and L. P. Heck, "Study on the effect of lexical mismatch in text-dependent speaker verification," in *Proc. Odyssey Speaker Recognition Workshop*, 2004.

[151] J. Pelecanos, U. Chaudhari, and G. Ramaswamy, "Compensation of utterance length for speaker verification," in *Odyssey*, Toledo, Spain, 2004.

[152] J. Naik and G. Doddington, "High performance speaker verification using principal spectral components," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Tokyo, Japan, 1986, pp. 881-884.

[153] H. Noda, K. Harada, and E. Kawaguchi, "A context-dependent Sequential decision for speaker verification," *IEICE Transactions on Information and Systems*, vol. E82-D, pp. 1433-1436, 1999.

[154] L. Rodriguez-Linares and C. Garcia-Mateo, "A novel technique for the combination of utterance and speaker verification systems in a text-dependent speaker verification task," in *Proc. International Conference on Speech Language Processing*, Sydney, Australia, 1998.

[155] W. Lit Ping and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 457-460 vol.1.

[156] J. R. Saeta and J. Hernando, "On the use of score pruning in speaker verification for speaker dependent threshold estimation," in *The Speaker and Language Recogition Workshop (Odyssey)*, Toledo, 2004, pp. 215-218.

[157] L. Allano, A. C. Morris, H. Sellahewa, S. Garcia-Salicetti, J. Koreman, S. Jassim, B. Ly-Van, D. Wu, and B. Dorizzi, "Non intrusive multi-biometrics on a mobile device:

a comparison of fusion techniques," in *Proc. SPIE Conference on Biometric Techniques for Human Identification III*, Orlando, 2006.

[158] A. Subramanya, Z. Zhengyou, A. C. Surendran, P. Nguyen, M. Narasimhan, and A. Acero, "A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, pp. IV-225-IV-228.

[159] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, 1995.

[160] T. Matsui and S. Furui, "Similarity normalization method for speaker verification based on a posteriori probability," in *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1994, pp. 59-62.

[161] M. Hebert and L. P. Heck, "Phonetic class-based speaker verification," in *EUROSPEECH*, 2003, pp. 1665-1668.

[162] M. Hebert and D. Boies, "T-norm for text-dependent commercial speaker verification applications: Effect of lexical mismatch," in *Proc. of ICASSP*, 2005, pp. 729–732.

[163] E. Mandler and J. Schurmann, "Combining the classification results of independent classifiers based on the Dempster/Shafer theory of evidence," in *Pattern Recognition and Artificial Intelligence*, North Holland, 1988, pp. 381-393.

[164] A. E. Hannani and D. Petrovska-Delacrétaz, "Fusing Acoustic, Phonetic and Data-Driven Systems for Text-Independent Speaker Verification," in *INTERSPEECH*, 2007, pp. 1240-1233.

[165] D. Burton, "Text-dependent speaker verification using vector quantization source coding," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 133-143, 1987.

[166] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Communication*, vol. 17, pp. 109-116, 1995.

[167] A. E. Rosenberg, O. Siohan, and S. Parathasarathy, "Speaker verification using minimum verification error training," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 105-108 vol.1.

[168] J. B. Pierrot, J. Lindberg, J. Koolwaaij, H. P. Hutter, D. Genoud, M. Blomberg, and F. Bimbot, "A comparison of a priori threshold setting procedures for speaker verification in the CAVE project," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 125-128 vol.1.

[169] N. Poh and S. Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognition*, vol. 39, pp. 223-233, 2006.

[170] K. Wadhwa, "Voice verification: technology overview and accuracy testing results," in *Biometrics*, International Biometric Group, London, UK, 2004.

[171] L. Yee Wah, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004, pp. 145-148.

[172] D. Matrouf, J. F. Bonastre, and C. Fredouille, "Effect of Speech Transformation on Impostor Acceptance," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. I-I.

[173] J. P. Campbell Jr and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proc. IEEE ICASSP*, Phoenix, 1999, pp. 829-832.

[174] R. Cole, M. Noel, and V. Noel, "The CSLU speaker recognition corpus," in *proceedings of ICSLP*, Sydney, 1998, pp. 3167-3170.

[175] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *INTERSPEECH*, Jeju Island, Korea, 2004, pp. 2489–2492.

[176] R. Cole, M. Noel , and V. Noel "The CSLU speaker recognition corpus," in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 3167-3170.

[177] D. Mazzoni and R. Dannenberg, "Audacity [software]," Pittsburg, PA: Carnegie Mellon University, 2000.

[178] H. Duxans, "Voice conversion applied to text-to-speech systems," PhD thesis, Universitat Politecnica de Catalunya, Barcelona, Spain, 2006.

[179] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 346-348 vol. 1.

[180] D. Sündermann, "Voice conversion Matlab toolbox," Technical Report, Siemens Corporate Technology, Munich, Germany2007.

[181] D. Sündermann and H. Ney, "An automatic segmentation and mapping approach for voice conversion parameter training," *Proc. of the AST'03, Maribor, Slovenia*, 2003.

[182] D. Sundermann, G. Strecha, A. Bonafonte, H. Höge, and H. Ney, "Evaluation of VTLN-Based Voice Conversion for Embedded Speech Synthesis," in *European Conference on Speech Communication and Technology*, 2005.

[183] D. Sundermann, A. Bonafonte, H. Ney, and H. Hoge, "Time Domain Vocal Tract Length Normalization," in *Proc. of the ISSPIT*, Italy, 2004, pp. 191-194.

[184] K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostin, F. Cardinaux, S. Marcel, S. Bengio, C. Sanderson, and N. Poh, "Face authentication test on the BANCA database," in *Proceedings of 17th International Conference on Pattern Recognition*, 2004, pp. 523-532.

[185] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 947-954.

[186] R. A. Dara, "Cooperative training in multiple classifier systems," Doctor of Philosophy, Waterloo, Ontario, Canada, 2007.

[187] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.

[188] J. Czyz, M. Sadeghi, J. Kittler, and L. Vandendorpe, "Decision fusion for face authentication," *Biometric Authentication*, pp. 1-6, 2004.

[189] N. Poh, T. Bourlai, J. Kittler, L. Allano, F. Alonso-Fernandez, O. Ambekar, J. Baker, B. Dorizzi, O. Fatukasi, and J. Fierrez, "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *IEEE Transactions on Information Forensics and Security,* vol. 4, pp. 849-866, 2009.

[190] N. Poh and S. Bengio, "How do correlation and variance of base-experts affect fusion in biometric authentication tasks?," *IEEE Transactions on Signal Processing,* vol. 53, pp. 4384-4396, 2005.

[191] L. Valet, G. Mauris, and P. Bolon, "A statistical overview of recent literature in information fusion," in *IEEE AESS Systems Magazine*, 2001, pp. 7-14.

[192] N. Poh and S. Bengio, "EER of Fixed and Trainable Fusion Classifiers: A Theoretical Study with Application to Biometric Authentication Tasks " in *Multiple Classifier Systems*. vol. 3541, N. Oza, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2005, pp. 19-25.

[193] A. R. Webb and K. D. Copsey, *Statistical pattern recognition*: John Wiley & Sons Inc, 2011.

[194] G. P. McCabe and D. S. Moore, *Introduction to the Practice of Statistics Ise*, 5 ed. New York: W.H.Freeman & Company, 2005.

[195] K. Venkataramani and B. V. K. Vijaya Kumar, "Role of statistical dependence between classifier scores in determining the best decision fusion rule for improved biometric verification," in *International Workshop on Multimedia Content Representation, Classification and Security (MRCS)*, 2006, pp. 489-496.

[196] E. N. Zois and V. Anastassopoulos, "Fusion of correlated decisions for writer verification," *Pattern Recognition,* vol. 34, pp. 47-61, 2001.

[197] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language,* vol. 22, pp. 17-38, 2008.

[198] E. Mandler and J. Schürmann, "Combining the classification results of independent classifiers based on the Dempster/Shafer theory of evidence," *Pattern recognition and artificial intelligence,* vol. 10, pp. 381–393, 1988.

[199] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin, "Is independence good for combining classifiers?," in *Proceedings of 15th International Conference on Pattern Recognition*, 2000, pp. 168-171 vol.2.

[200] S. Tulyakov and V. Govindaraju, "Utilizing Independence of Multimodal Biometric Matchers," in *Multimedia Content Representation, Classification and Security*. vol. 4105, B. Gunsel, *et al.*, Eds., ed: Springer Berlin / Heidelberg, 2006, pp. 34-41.

[201] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Machine Learning,* vol. 29, p. 103, 1997.

[202] O. Ushmaev and S. Novikov, "Biometric Fusion: Robust Approach," in *Proceedings of 2nd Workshop on Multimodal User Authentication (MMUA)*, Toulouse, France, 2006.

[203] K. M. Ali and M. J. Pazzani, "On the link between error correlation and error reduction in decision tree ensembles," Technical Report 95-38, Department of Information and Computer Science, University of California, Irvine, 1995.

[204] S. Nakagawa, "Correlation Analysis of Speaker Differences in Vowels, Consonants and Spoken Digits," *Studia phonologica,* vol. 21, pp. 90-99, 1987.

[205] P. Niyogi, "Modelling speaker variability and imposing speaker constraints in phonetic classification," Master's Thesis, Massachusetts Institute of Technology, 1991.

[206] C. Ji and S. Ma, "Combinations of weak classifiers," *IEEE Transactions on Neural Networks,* vol. 8, pp. 32-42, 1997.

[207] M. Gal-Or, J. H. May, and W. E. Spangler, "Assessing the predictive accuracy of diversity measures with domain-dependent, asymmetric misclassification costs," *Information Fusion,* vol. 6, pp. 37-48, 2005.

[208] S. Bian and W. Wang, "On diversity and accuracy of homogeneous and heterogeneous ensembles," *International Journal of Hybrid Intelligent Systems,* vol. 4, pp. 103-128, 2007.

[209] M. Kam, Q. Zhu, and W. S. Gray, "Optimal data fusion of correlated local decisions in multiple sensor detection systems," *IEEE Transactions on Aerospace and Electronic Systems,* vol. 28, pp. 916-920, 1992.

[210] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers," *Connection Science,* vol. 8(3-4), pp. 385–403, 1996.

[211] R. A. Jacobs, "Methods for combining experts' probability assessments," *Neural computation,* vol. 7, pp. 867-888, 1995.

[212] E. N. Zois and V. Anastassopoulos, "Fusion of correlated decisions for writer identification," in *Handwriting Analysis and Recognition (Ref. No. 1998/440), IEE Third European Workshop on*, 1998, pp. 3/1-3/7.

[213] J. Kittler, J. Matas, K. Jonsson, and M. Ramos Sánchez, "Combining evidence in personal identity verification systems," *Pattern Recognition Letters,* vol. 18, pp. 845-852, 1997.

[214] G. Salton, C. Buckley, and C. Yu, "An evaluation of term dependence models in information retrieval," *Research and Development in Information Retrieval,* pp. 151-173, 1983.

[215] R. M. Losee, "Term dependence: truncating the Bahadur Lazarsfeld expansion," *Information processing & management,* vol. 30, pp. 293-303, 1994.

[216] C. T. Yu, C. Buckley, K. Lam, and G. Salton, "A generalized term dependence model in information retrieval," *Information Technology: Research and Development,* vol. 2(4), pp. 129-154, 1983.

[217] L. I. Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 32, pp. 146-156, 2002.

[218] G. Giacinto and F. Roli, "Dynamic classifier selection," *Multiple Classifier Systems,* pp. 177-189, 2000.

[219] J. Kittler and F. Roli, "Multiple classifier systems." vol. 1857, J. Kittler and F. Roli, Eds., ed: Springer-Verlag Pub., Lecture Notes in Computer Science, 2000, pp. 1-404.

[220] G. Giacinto and F. Roli, "Dynamic classifier selection based on multiple classifier behaviour," *Pattern Recognition,* vol. 34, pp. 1879-1882, 2001.

[221] J. Cao, M. Ahmadi, and M. Shridhar, "Recognition of handwritten numerals with multiple feature and multistage classifier," *Pattern Recognition,* vol. 28, pp. 153-160, 1995.

[222] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, and A. Gelzinis, "Soft combination of neural classifiers: A comparative study," *Pattern Recognition Letters,* vol. 20, pp. 429-444, 1999.

[223] K. Woods, W. P. Kegelmeyer Jr, and K. Bowyer, "Combination of multiple classifiers using local accuracy estimates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, pp. 405-410, 1997.

[224] L. I. Kuncheva, "Clustering-and-selection method for classifier combination," in *Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies*, 2000, pp. 185-188

[225] R. Liu and B. Yuan, "Multiple classifiers combination by clustering and selection," *Information Fusion,* vol. 2, pp. 163-168, 2001.

[226] T. Ho, "Complexity of classification problems and comparative advantages of combined classifiers," in *Multiple Classifier Systems*. vol. 1857, J. Kittler and F. Roli, Eds., ed: Springer-Verlag, LNCS, 2000, pp. 97-106.

[227] A. Sharkey, N. Sharkey, U. Gerecke, and G. Chandroth, "The “test and select” approach to ensemble combination," *Multiple Classifier Systems,* pp. 30-44, 2000.

[228] H. Altinçay, "Comparing Diversity and Training Accuracy in Classifier Selection for Plurality Voting Based Fusion," in *Adaptive and Natural Computing Algorithms*, B. Ribeiro, *et al.*, Eds., ed: Springer Vienna, 2005, pp. 381-384.

[229] J. Kittler, "Feature set search algorithms," in *Pattern recognition and signal processing*. vol. 41, C. H. Chen, Ed., ed: Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, 1978, p. 60.

[230] D. Magee, "A sequential scheduling approach to combining multiple object classifiers using cross-entropy," *Multiple Classifier Systems,* pp. 161-161, 2003.

[231] R. E. Schapire, "The boosting approach to machine learning: An overview," *Lecture Notes in Statistics,* pp. 149-172, 2003.

[232] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceeding of IEEE Computer Vision and Pattern Recognition*, 2001, pp. I-511-I-518 vol. 1.

[233] A. J. C. Sharkey and N. E. Sharkey, "Combining diverse neural nets," *The Knowledge Engineering Review*, vol. 12, pp. 231-247, 1997.

[234] D. Ruta and B. Gabrys, "Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems," in *Proceedings of the 4th International Symposium on Soft Computing*, 2001, pp. 1824–025.

[235] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*: Wiley Inter-science, 2004.

[236] K. F. Goebel and W. Yan, "Using correlation-based measures to select classifiers for decision fusion," in *Proceedings of SPIE*, 2005, pp. 180-191.

[237] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing,* vol. 19, pp. 699-707, 2001.

[238] F. Roli, G. Giacinto, and G. Vernazza, "Methods for Designing Multiple Classifier Systems," in *Multiple Classifier Systems*. vol. 2096, J. Kittler and F. Roli, Eds., ed: Springer Berlin / Heidelberg, 2001, pp. 78-87.

[239] C. A. Shipp and L. I. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion,* vol. 3, pp. 135-148, 2002.

[240] T. Lofstrom, U. Johansson, and H. Bostrom, "On the Use of Accuracy and Diversity Measures for Evaluating and Selecting Ensembles of Classifiers," in *Seventh International Conference on Machine Learning and Applications,* 2008, pp. 127-132.

[241] M. Aksela and J. Laaksonen, "Using diversity of errors for selecting members of a committee classifier," *Pattern Recognition,* vol. 39, pp. 608-623, 2006.

[242] D. Partridge and W. Krzanowski, "Distinct failure diversity in multiversion software," *Research report,* vol. 348, 1997.

[243] S. Wang and X. Yao, "Relationships Between Diversity of Classification Ensembles and Single-Class Performance Measures," *IEEE Transactions on Knowledge and Data Engineering,*, pp. 1-1, 2011.

[244] S. W. Looney, "A statistical technique for comparing the accuracies of several classifiers," *Pattern Recognition Letters,* vol. 8, pp. 5-9, 1988.

[245] M. Goldstein, "An Appropriate Test For Comparative Discriminatory Power," *Multivariate Behavioral Research,* vol. 11, pp. 157-163, 1976.

[246] G. Tsoumakas, L. Angelis, and I. Vlahavas, "Selective fusion of heterogeneous classifiers," *Intelligent Data Analysis,* vol. 9, pp. 511-525, 2005.

[247] H. Bozdogan, "Multi-sample cluster analysis as an alternative to multiple comparison procedures," *Bulletin of Informatics and Cybernetics,* vol. 22, pp. 95-130, 1986.

[248] L. Madden, J. Knoke, and R. Louie, "Considerations for the use of multiple comparison procedures in phytopathological investigations," *Phytopathology,* vol. 72, pp. 1015-1017, 1982.

[249] C. Gates and J. Bilbro, "Illustration of a cluster analysis method for mean separation," *Agronomy Journal,* vol. 70, pp. 462-465, 1978.

[250] K. Veeramachaneni, L. A. Osadciw, and P. K. Varshney, "An adaptive multimodal biometric management algorithm," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* vol. 35, pp. 344-356, 2005.

[251] A. Kumar, V. Kanhangad, and D. Zhang, "A new framework for adaptive multimodal biometrics management," *IEEE Transactions on Information Forensics and Security,* vol. 5, pp. 92-102, 2010.

[252] D. Suendermann, "Text Independent Voice Conversion," PhD thesis, Bundeswehr University Munich, 2007.